


This PDF file contains all parts of


GBIF Training Manual 1: Digitisation of Natural History Collections

This book contains many hyperlinks. Most are to websites (if you are using the document offline, you will not be able to follow these).


Some of the links are internal to the document (particularly in the Introduction).


➔ To go to a different place in the document

 Mouse over a link and click to jump to the part of the book that is of interest, or

 Click on the Bookmark in the bookmarks sidebar

➔ To return to your previous place, either

 Click on its Bookmark in the bookmarks sidebar, or

 Hit Alt + left arrow (or right click and select “previous view”)

[START HERE](#)



**GBIF TRAINING
MANUAL 1:**

**DIGITISATION OF
NATURAL
HISTORY
COLLECTIONS
DATA**

GBIF Training Manual 1: Digitisation of Natural History Collections Data

Published by:
Global Biodiversity Information Facility
<http://www.gbif.org>

© 2008, Global Biodiversity Information Facility

Digitisation of Natural History Collections Data ISBN: 87-92020-07-0

Permission to copy and/or distribute all or part of the information contained herein is granted, provided that such copies carry due attribution to the Global Biodiversity Information Facility (GBIF).

Recommended citation:

Global Biodiversity Information Facility. 2008. *GBIF Training Manual 1: Digitisation of History Collections Data*, version 1.0. Copenhagen: Global Biodiversity Information Facility.

While the editor, authors and the publisher have attempted to make this book as accurate and as thorough as possible, the information contained herein is provided on an "As Is" basis, and without any warranties with respect to its accuracy or completeness. The editor, authors and the publisher shall have no liability to any person or entity for any loss or damage caused by using the information provided in this book.

GBIF Training Manual 1

Digitisation of Natural History Collections Data

Covers and copyright page	Covers and copyright
Introduction	Introduction
Meredith A. Lane	
Chapter 1: Uses of Digitised Collections Data	Chapter 1
Arthur D. Chapman	
Chapter 2: Initiating a Collection Digitisation Project	Chapter 2
Christopher K. Frazier, John Wall and Sharon Grant	
Chapter 3: Data Quality	Chapter 3
Arthur D. Chapman	
Chapter 4: Data Cleaning	Chapter 4
Arthur D. Chapman	
Chapter 5: Georeferencing	Chapter 5
Arthur D. Chapman and John Wieczorek (eds.)	
Chapter 6: Generalising Sensitive Data.....	Chapter 6
Arthur D. Chapman and Oliver Grafton	
Glossary and Acronym Expansion	Glossary and Acronym Expansion
GBIF Secretariat staff	
	booklet
Appendix: GBIF Data Portal Tutorial	
Donald Hobern and Meredith A. Lane	online (English) online (French)
(French translation by GBIF-fr)	

Introduction

The Global Biodiversity Information Facility ([GBIF](#)) is a worldwide network that makes primary, scientific, biodiversity data (documented species occurrence data) from many sources openly available via the Internet. It does this by building an information infrastructure that interconnects hundreds of databases, and by promoting the digitisation and sharing of data that are not currently available via the Internet, such as those associated with specimens in natural history museums.

This promotion of digitisation is approached in a number of ways:

- Seed money awards to stimulate digitisation projects;
- The development (with partners) of community-accepted standards for data and metadata, as well as software tools that enable interconnectivity and interoperability;
- Workshops for training in digitisation and data-sharing; and
- Guides such as this Training Manual and its components.

GBIF's hope is to help collections and database personnel around the world share best practices in the tasks and operations required in building a web-based, global "natural history collection and herbarium" that can be accessed any time any where by any one via the Internet.

GBIF is based upon primary scientific data—data that were recorded directly from nature—and upon a robust and comprehensive taxonomic system. These kinds of data can be used and reused in different analyses without diminishing their value.

However, in this digital age, the use of biodiversity data are limited by the paucity of records that are in a digital form; most data are recorded only on paper in ink. For this reason, GBIF places a strong emphasis on the digitisation of natural history and other biological collections, as well as taxonomic names data and concepts. In addition, GBIF will provide tools that will, to unprecedented levels, enhance quality of these data, and describe its fitness for various uses.

GBIF comprises its Participants, their Nodes, data providers around the world, and a coordinating Secretariat that works with partner organisations of many types to accomplish the goals of all. It does this by

- Supporting and promoting the view that sharing biodiversity data, with clear rules and with *full respect for the rights of the providers*, has clear advantages for both users and providers.
- Reaching out to data providers and potential users of the data, *providing them with opportunities to increase their capacity to share* and utilise biodiversity data.
- *Encouraging and facilitating the digitisation of data*, including historical specimens, their label texts and associated materials, as well as observational data, so that these can be added to the digital store of available data;
- Encouraging and facilitating the digital capture, documentation and georeferencing of newly gathered specimens and observational records;
- Building an information architecture that offers web services to users and data providers, and makes biodiversity databases interoperable among themselves and across levels of biological organisation, as well as with digital literature.

The founding [GBIF Memorandum of Understanding](#) laid out the principles of GBIF, to which it still adheres. These include that GBIF will:

- be shared and distributed, while encouraging co-operation and coherence;
- be global in scale, though implemented nationally, regionally and locally;
- be accessible by individuals anywhere in the world, offering potential benefits to all, while being funded primarily by those that have the greatest financial capabilities;
- promote standards and software tools designed to facilitate their adaptation into multiple languages, character sets and computer encodings;
- disseminate technological capacity by making widely available scientific and technical information; and
- make biodiversity data universally available, *while fully acknowledging the contribution made by those gathering and furnishing these data.*

GBIF's adherence to these principles is specifically intended to achieve benefits both for the users and providers of primary species occurrence data, and to make GBIF a global public good.

Benefits of GBIF

To users: GBIF is a Global Public Good

Public goods, as generally understood, have two important characteristics: (1) they are *freely available to all*, and (2) *they are not diminished by use*. By this definition alone, GBIF is a public good: (1) GBIF's fundamental principle is freely shared, accessible data, and (2) GBIF-mediated data can be used and reused by anyone—the very use of the data can often improve their quality. There is no single major “buyer” for GBIF-mediated data – in fact, the data can be used by researchers to generate new knowledge, by non-governmental agencies of one sort or another, or by governments for decision-making (among many other uses, see Chapter 1 of this *Manual*).

A recent paper by Arzberger *et al.* (2004) has as its core principle that publicly funded research data should be openly available to the maximum extent possible because these data are a public good produced in the public interest. A set of very good examples of this are the databanks such as GenBank, PDB and FlyBase. They are supported by public funds, and are used for free by basic researchers to generate new knowledge as well as by the private sector to generate profits. Similarly, GBIF is a databank (though unlike these others, it is a *distributed* one) for the species level of biodiversity – serving up data from many sources that were generated in large part using public funds.

Benefits to data providers

Frazier and colleagues (see Chapter 2 of this *Manual*) provide a number of reasons that holders of species level biodiversity data should share those data, among which are:

- wider dissemination, thereby raising the profile of the institution;
- facilitating research through reducing transcription time and enabling novel combinations of species data with other data types;
- enhancing curatorial activities; and
- importantly, protecting the specimens by reducing handling and shipment.

Other good reasons are provided in by Townsend & Navarro (2002) and Townsend, et al. (2005).

One of the reasons for digitisation that GBIF emphasises to potential data providers is that by digitising and sharing data, those data that would otherwise sit in jars or on sheets behind closed cabinet doors are actually *used*. There is no value to data that are not used – and because the arguments for continued support of natural history collections always contain statements about how valuable the associated data are, digitisation and sharing make those arguments ever more valid. There is a value *to* the digitisation and sharing of data as well (see [Chapter 1](#) of this *Manual*) – as shared data are used, the users can help with cleaning and quality improvement. As the quality increases, so does usage – and arguments for continued support of the collections that underlie and vouch for those data gain even more strength.

In addition, for data providers, GBIF makes available a convenient means of joining the global community – the shared standards and protocols that emerge from GBIF’s consultative processes can be adopted by anyone, both those involved in the development of the standards and those who wish to share data but are not inclined to participate in IT development. This is in stark contrast to the situation only 10 to 15 years ago, when each collection that digitised was an idiosyncratic silo, built by small groups in isolation. Today’s world of biodiversity informatics has changed all that.

GBIF truly hopes that the community finds this *Training Manual* welcome and useful, and in turn welcomes feedback and improvements on the text. All components of this *Manual*, and its planned future companion volumes, are dynamic and community-owned documents that can be upgraded and updated.

Open Access to Data, and the Free Exchange of Information

GBIF was founded on principles of open availability of data (Memorandum of Understanding, [Paragraph 8](#)): *To the greatest extent possible, GBIF is an open-access facility. All users, whether GBIF Participants or others, should have equal access to data in databases affiliated with or developed by GBIF.* Indeed, the Governing Board has affirmed this principle in a [Recommendation](#) to research councils, other funding agencies and private foundations that they promote the

1. *maintenance and sharing of digital biodiversity data generated in projects funded by them, and that*
2. *these data be made publicly available, within a specified time following the completion of the research, through mechanisms that cooperate with GBIF.*

In addition, GBIF has [recommended](#), as a contribution to the goals of [Article 17 of the Convention on Biological Diversity](#) on the Exchange of Information, that

1. *natural history institutions that house biodiversity materials from other countries ensure that species- and specimen-level data and metadata be digitised and made openly and publicly available through mechanisms that cooperate with GBIF, and that*
2. *research organisations and councils, governmental and non-governmental organisations, national and international funding agencies, and private foundations provide funding for ... activities that include the digitisation and open dissemination of species- and specimen-level data, in accordance with GBIF-mediated standards and protocols.*

Special considerations related to sensitive data (for example, those related to threatened species, are covered in [Chapter 6](#) of this *Manual*.

GBIF and Intellectual Property Rights (IPR)

From its very beginning, GBIF has been concerned about the IPR of those persons, institutions and organisations who share data through its network. This is so fundamental to the organisation that [Paragraph 8](#) of the founding document, the GBIF [Memorandum of Understanding](#), is devoted to IPR.

In addition, GBIF has established a *pro bono* Legal Advisory Group (proLEG) to provide advice and recommendations to GBIF, its users, and its data providers concerning IPR issues. In the first [proLEG report](#) to GBIF, this group made several recommendations, which included:

1. *Considering that the mandate and purpose of GBIF is to promote the sharing of primary biodiversity data freely and openly, GBIF should seek to rely upon and use, as much as possible, the practices, norms, and policies of public science to guide its activities and **avoid using legalistic solutions and enforcement mechanisms.***
2. *Consistent with Recommendation 1 and the relevant statutory law, GBIF should **impose the least possible restrictions and obligations on users.***
3. *GBIF should **continue to include attribution as a condition of the use of the data** through its portal in order to encourage such normative behavior by the data users.*
4. *GBIF should continue to **work with its data providers to promote its free and open data access policy**, subject only to appropriate attribution. For those data providers that require restrictions on commercial reuse of their data, the development of a standardized licensing mechanism similar to the Creative Commons licenses could be appropriate.*

GBIF continues to engage in consultative activities with proLEG as well as other international organisations (e.g. [CODATA](#)) with the aim of providing general guidelines in these and other IPR-related areas. At the same time, GBIF does have in place both a Data Sharing Agreement and a Data Use Agreement, both of which follow the recommendations above.

[GBIF Data Sharing Agreement](#)

When a data provider registers a dataset with the GBIF UDDI registry, they agree to abide by the provisions of the GBIF Data Sharing Agreement, which has been formulated with guidance from proLEG.

[GBIF Data Use Agreement](#)

Likewise, any time a user accesses data that are shared across the GBIF network, they are first required to agree to the GBIF Data Use Agreement, which has also been formulated with guidance from proLEG.

More on the GBIF Data Policy, and its stance on Open Access, can be found in the *Pamphlets* section of http://www.gbif.org/GBIF_org/GBIF_Documents.

References

Arzberger, P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhler and P. Wouters. 2004. Promoting access to public research data for scientific, economic and social development. [Data Science Journal 3: 135 – 152.](#)

Peterson, A.T. and Navarro-Sigüenza, A.G. 2002. Computerising Bird Collections and Sharing Data Openly: Why bother? [Bonner Zoologische Beiträge 51: 205-212.](#)

Peterson, A. T., C. Cicero and J. Wieczorek. 2005. Free and open access to bird specimen data: Why? [The Auk 122: 987-990.](#)

Uses of Digitised Collections Data

Introduction to Chapter 1	2
Taxonomy	7
Biogeographic Studies	14
Species Diversity and Populations	18
Life Histories and Phenologies	23
Endangered, Migratory and Invasive Species	24
Impact of Climate Change	30
Ecology, Evolution and Genetics	32
Environmental Regionalisation	37
Conservation Planning	39
Natural Resource Management	43
Agriculture, Forestry, Fisheries and Mining	44
Health and Public Safety	53
Bioprospecting	56
Forensics	57
Border Control and Wildlife Trade	59
Education and Public Outreach	61
Ecotourism	65
Art and History	67
Society and Politics	69
Recreational Activities	71
Human Infrastructure Planning	73
Aquatic and Marine Biodiversity	75
Conclusion	75
References	76
Index to Chapter 1	88

This Chapter is equivalent to:

Chapman, A. 2005. *Uses of Primary Species Occurrence Data*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 106 pp. ISBN: 87-92020-01-1 (available as a standalone PDF from <http://www.gbif.org>)

Introduction to Chapter 1

Plant and animal specimen data held in museums and herbaria, survey data and species observational data provide a vast information resource, providing not only present day information on the locations of these entities, but also historic information going back several hundred years (Chapman and Busby 1994). It is estimated that there are approximately 2.5-3 billion collections worldwide in museums, herbaria and other collection institutions (Duckworth *et al.* 1993, OECD 1999). In addition there are untold numbers of observational data records. Projects to digitise this information are underway in many institutions, with others at either the discussion or planning stage.

A key purpose of digital information in the biological sciences is to provide users of information with a cost-effective method of querying and analysing that information. The biological world is infinitely complex and must be generalised, approximated and abstracted in order to be represented and understood (Goodchild *et al.* 1991). Ways of presenting biodiversity information to users is through the use of geographic information systems, environmental modelling tools, decision support systems, books, CDs, images and on-line databases, specimens and their parts, DNA reports, etc. Within these tools, however, it is essential that variation be sampled and measured, and error and uncertainty be described and visualised. It is in this area that we still have a long way to go (Goodchild *et al.* 1991).

The uses of primary species-occurrence data are wide and varied and encompass virtually every aspect of human endeavour – food, shelter and recreation; art and history, society, science and politics. The examples shown in this paper emphasizes the importance of having museum specimen data digitized and made available to the wider user community. In this way, the collections will be made even more valuable than they already are, and provide new opportunities for funding and collaboration through their increased relevance and value to a much larger audience. With dwindling resources being made available for the biological sciences, funding bodies are beginning to ask the relevance of many natural history collections, and it is becoming increasingly more difficult to obtain funds for collection maintenance. By making information available to the broader scientific community for use in conservation and the many other areas of study covered in this paper, institutions will have a much more robust and sustainable argument for continued funding. In addition, it will rapidly add to the world's knowledge of biodiversity and ecological systems and aid in its future conservation and sustainable use and management.

The increased availability of data on species is opening up new and improved methods of dealing with these issues. The information in museums is a storehouse going back hundreds of years, and the new availability of that storehouse in on-line databases is improving science, reducing costs by providing for more efficient and effective biological survey, freeing up scientists to spend more time on research, and leading to a more rapid build-up of knowledge of our environments leading to its improved conservation and sustainable use.

Taxonomic research is benefiting through the availability of images of specimens, including types, data on the location of specimens in other museums, etc. But perhaps the greatest benefit of the availability of distributed data is the study of the biogeography of species – their location in time and space. “By reducing the costs of studying vectors of human disease, biological invasions, and global climate change, biological collections provide direct financial and social benefits to society” (Suarez and Tsutsui 2004).

One of the things that will come out of a study of uses of species-occurrence data is the opening up new requirements for recording information as part of future collecting events (Chapman 2005b). This may even include a greater use of digital images (Basset *et al.* 2000), and video. But along with all the positives of electronic data exchange, there is a tendency to divorce the data from the objects, and it is important that those outside the museum community recognise that the objects themselves remain important long-term repositories and sources of data that have yet to be captured and developed (Winker 2004). Ultimately, maintaining and developing the infrastructure of biodiversity collections will produce unforeseen benefits (Suarez and Tsutsui 2004). Those benefits to society will be multiplied through the ready availability of the information to those that need to use them.

But primary species-occurrence data are not just the data held in museums and herbaria. There is a massive amount of observational and survey data held in universities, by non-governmental organisations and by private individuals and these data add valuable additional knowledge on our environment. They are not competing data resources but complementary and each have their strengths and weaknesses in supplying the information the world needs.

Some question the value of digitised museum specimen data for use in biogeographic and other studies because much of the data are “outdated and unreliable”, with many records misidentified or badly geo-referenced (Wheeler *et al.* 2004). That may be true for many records, but, as shown in this paper, there are many other records that are not so unreliable, and that are being used by researchers and others with great success. The museum community is aware of the problems inherent in their data and are making concerted attempts to improve the quality of those data (Chapman 2005a), and as stated by Edwards (2004), “one of the best ways to expose those errors is to make the data visible, so that qualified researchers can compare and correct them”. All data have errors, but that should not be a reason not to use the data, but to ensure that the error is documented and that users are made aware of the errors so that they may determine the fitness for use of the data (Chapman 2005b).

There are many uses for primary species data. Traditionally, collections in museums and herbaria were only made with one main purpose in mind – that of taxonomic study. Their long-term mission, however, is to document biodiversity and its distribution through time and space for research and education (Winker 2004) and to serve the public. The introduction of computer processing and computer databases have opened up this vast data store to many new uses (Chapman 1999). These uses include biogeographic studies (Longmore 1986, Peterson *et al.* 1998), conservation planning (Faith *et al.* 2001), reserve selection (Margules and Pressey 2000), development of environmental regionalisations (Thackway and Cresswell 1995), climate change studies (Chapman and Milne 1998, Pouliquen and Newman 1999, Peterson *et al.* 2002a), agriculture, forestry and fishery production (Booth 1996, Nicholls 1997, Cunningham *et al.* 2001), species translocation studies (Panetta and Mitchell 1991, Soberón *et al.* 2000, Peterson and Veiglas 2001), etc., etc. These and other uses will be elaborated further in this document. Many of these studies have used environmental modelling using software such as BIOCLIM (Nix 1986, Busby 1991), GARP (Stockwell and Peters 1999, Pereira 2002) or methods such as Generalised Linear Models (GLM) (Austin 2002). Most of these species distribution models rely on specimen or observation records, generally of a presence-only nature (usually including records from herbaria or museums as well as observation data) or occasionally presence-absence data from systematic surveys.

Much of the data (both museum and observational) have been collected opportunistically rather than systematically (Chapman 1999, Williams *et al.* 2002) and this can result in large spatial biases – for example, collections that are highly correlated with road or river networks

(Margules and Redhead 1995, Chapman 1999, Peterson *et al.* 2002, Lampe and Riede 2002). Museum and herbarium data and most observational data, generally only supply information on the presence of the entity at a particular time and says nothing about absences in any other place or time (Peterson *et al.* 1998). This restricts their use in some environmental models, but they remain the largest and most complete database of biological information over the last 200+ years we are ever likely to have. The cost of replacing these data with new surveys would be prohibitive. It is not unusual for a single survey to exceed \$1 million to conduct (Burbidge 1991). Further, because of their collection over time, they provide irreplaceable baseline data about biological diversity during a time when humans have had tremendous impact on such diversity. They are an essential resource in any effort to conserve the environment, as they provide the only fully documented record of the occurrence of species in areas that may have undergone habitat change due to clearing for agriculture, urbanization, climate change, or been modified in some other way (Chapman 1999).

But primary species data do not stop with just the information on the label, as there is information contained within the collections themselves and this may be used for tissue sampling, chemical analysis of contaminants, forensic information held in the DNA of individual specimens, etc. Living culture collections of micro-organisms that cannot otherwise be preserved, images and even video of individual birds and animals in the field, of preserved specimens in museums, or micrographs of parts, and even drawn illustrations – some done before photography was invented – must also be regarded as an integral part of the species-occurrence data record.

Data interchange and distributed data

As early as 1974, discussions on developing standards for electronic exchange of primary specimen data between museums and herbaria were taking place. Although the Internet was restricted to users in a limited research community and not generally available to biodiversity institutions (Kristula 2001), and exchange via media such as floppy disks, and magnetic tape was occurring around the world, no standards for doing so existed. As a result of these discussions, a standard for the interchange of biotaxonomic information was developed in Australia in 1979 (Busby 1979). Later, the Australian herbaria got together and extended this standard for use by botanical institutions and the HISPID (Herbarium Information Standards for the Interchange of Data) standard was developed (Croft 1989, Conn 1996, 2000). Although very few institutions used these standards for interchange, many used them as a template for designing their databases. The HISPID standard was later adopted as a TDWG (Taxonomic Databases Working Group) standard.

The development of the Internet, and especially the World Wide Web (Berners-Lee 1999), allowed new opportunities for the interchange of data. Although the Environmental Resources Information Network (ERIN) used distributed data for modelling on the Internet as early as 1994 (Boston and Stockwell 1995), there were few other successful electronic data interchange projects that utilised the internet until the Species Analyst (Vieglas 1999, 2003a) project began the late 1990s.

Since then, a number of distributed projects have begun, including the Red Mundial de Información sobre Biodiversidad (REMIB) –The World Network on Biodiversity (CONABIO 2002), Australian Virtual Herbarium (CHAH 2002), *speciesLink* (CRIA 2002), European Natural History Specimen Information Network (ENHSIN) (Güntsch 2004), Biological Collection Access Service for Europe (BioCASE 2003), the Mammal Networked Information System (MaNIS 2001), and the GBIF Portal (GBIF 2004). These systems use on-line information retrieval to search databases maintained in the home institutions, extracting

data in a way similar to what Google does for web resources. Early versions of these relied on the information retrieval standard developed primarily for library use – Z39.50 (NISO 2002), but more recently the museums community have combined to develop new standards, the Darwin Core Schema (Vieglais 2003b) along with the DiGIR protocol (SourceForge 2004) and the combined BioCASE protocol (BioCASE 2003) and ABCD (Access to Biological Collections Data) schema (TDWG 2004) that are more fitted for interchange of primary species information. More recently the Taxonomic Databases Working Group and others have begun working to develop a combined protocol (TAPIR - <http://ww3.bgbm.org/tapir>) treading a middle path between the simplicity of DiGIR and the complexity of BioCASE.

Multiple uses

Most projects that use species-occurrence data incorporate more than just one type of use. As evident from this paper, there is considerable overlap in uses within any one project. A project might include mapped primary records, some taxonomic study (possibly involving the use of character databases), environmental modelling and predictive distributional studies which may involve endangered or migratory species, climate change impact studies as well as population viability analysis and studies of species associations, ecology and evolutionary history. The project may then involve species recovery studies and monitoring, as well as development of environment protection legislation, reserve and conservation assessment, links to border and custom controls to prevent illegal smuggling, and finally education and social links. It is sometimes difficult to identify where one use stops and another begins, and I hope readers will excuse the inevitable overlap that is evident throughout this paper.

The ability to search databases all around the world for spatially-referenced primary species-occurrence data has opened up the information to a range of uses, many of which have previously not been possible. This paper will elaborate on some of those uses and present examples. It should be noted that it is beyond the scope of a paper such as this to cover every example of use – examples given are just that – samples to illustrate the types of uses mentioned.

Some of this overlap in uses can be seen from the first GBIF Demonstration Project in 2003. (UTU-Biota 2004).

GBIF Demonstration Project 2003

The first GBIF Demonstration Project (<http://gbifdemo.utu.fi/>) provided a number of user-friendly examples of how primary biodiversity data can effectively be used, managed, exchanged and disseminated via the Internet. It was prepared for GBIF by the University of Turku in association with the Institute of Amazonian Research (IIAP). The project was divided into four sections or “tours”. Tour 1 dealt with Neotropical species distributions, Tour 2 with multi-authored rainforest trees inventories, Tour 3 with sub-arctic plant observations and Tour 4 on planning and management of biodiversity.

In 2004, GBIF funded two more Demonstration projects (<http://www.gbif.org>). The first of these is an Australian-based project to develop an internet-based tool for biogeographic analysis of endemism and taxonomic distinctness. The second project is based in Mexico, and will demonstrate the feasibility of estimating the rate of disappearance of species populations by estimating distribution areas of species associated with primary vegetation on the basis of primary biodiversity data. Both will use data extracted via the GBIF Portal.

Benefits of making species-occurrence data available

Many of the uses of species-occurrence data elaborated in this paper have required the user to visit the collections institution – the museum or herbarium, etc. to seek access to the information, or to obtain identifications. Staff of the museum then has to spend time and resources in identifying the material for the user (which may be from hundreds to thousands a year for some collectors (Suarez and Tsutsui 2004) or readying the data for the user. Huge resources are spent each year as scientists travel to museums to use the collections, or as museums loan specimens to researchers. Between 1976 and 1986, the Smithsonian's entomological collection loaned, on average, over 100,000 specimens each year (Miller 1991) and it, like most of the world's larger museums, annually hosts hundred of visiting researchers. Collections institutions are now beginning to realise that they can save valuable time and resources by making available electronically as much of that data as is possible. An example is with the Botanischer Garten und Botanisches Museum Berlin-Dahlem where their herbarium loan system has been completely replaced with a digital loan system¹ (<http://ww2.bgbm.org/Herbarium/AccessLoanNew.cfm>). Not only does it free up resources, more often than not, those resources are the taxonomists and researchers that can then spend more time on basic research and curation and less on administration and on helping others. The digitisation of the hundreds of millions of collections held in natural history museums, however, is no small task and will take many years, or even decades to complete.

The increased use of species data through distributed systems will provide a climate that will allow, among others:

- Consolidation of collections infrastructure and holdings within museums, herbariums, botanical gardens, zoological gardens, germplasm banks, etc.;
- A reassignment of resources toward increased research and curation;
- Improvements in the standardization, quality, maintenance and organization of important biodiversity collections;
- Reduce physical handling of specimens, ensuring their longevity;
- Reduce costs of shipping, insurance, etc. of transferring loans and specimens between institutions;
- The sharing of information between institutions and researchers, including with countries of origin;
- A more rapid advancement of the biodiversity knowledge-base as researchers build on the information in a more timely manner;
- Establishment of international biodiversity information networks between institutions involved with biodiversity research, conservation, genetics, production, resource management, tourism, etc.;
- Improvements in the management and availability of image, cartographic, genetic, and other databases that will subsidize biodiversity research;
- Improvements in the management of conservation units as knowledge about biodiversity becomes more readily available;
- Improved evaluation of the representativeness of existing conservation units and reserves, and the identification of priority areas for the establishment of new ones;
- Development of projects to study problems that affect conservation, such as the effects and consequences of habitat fragmentation and climate change on biodiversity;

¹ Pers. comm.. Anton Güntsch, BGBM 2005.

- Improvements in border controls for managing and monitoring movements in endangered species, pests and diseases as identification tools and knowledge about the distributions of taxa are improved;
- Production and dissemination of checklists of all known biota of conservation areas, regions, States, and countries, etc.;
- Increased and more efficient production of identification tools, keys, catalogues and monographs (electronic and/or paper publications);
- More and improved inventories and studies for identifying biodiversity information gaps (both taxonomic and geographic);
- Development of research projects that aim at understanding the temporal and spatial distribution of biological diversity processes and functions;
- Comparative and retrospective studies for estimating biodiversity loss within regions, habitats, ecosystems, and across political and geographic boundaries;
- Comparative studies on environmental impact, such as climate change, urbanization, agriculture, fisheries, etc. and establishment of reference patterns for evaluation and monitoring of environmental impact with respect to biological diversity;
- Increased opportunities for bioprospecting, and the linking of programs with related and similar interests;
- Improvements in capacity building in biodiversity and biodiversity-related subjects;
- The development of professionals in new fields of knowledge and in new interfaces, such as biodiversity informatics, image services, and geographic information systems;
- Production of improved teaching material, such as field guides, identification keys, image databases, and on-line information for students and educators;
- Improved guides and information resources for use in ecotourism;
- Improved rates of publishing in taxonomy as researchers spend less time on identifications and on making data available on an individual basis;
- Improved linkages with local people for collecting, ecological research and preliminary identification using parataxonomists;
- transfer of some of the burden of sorting and preliminary identification of field samples from the extremely small number of highly-skilled taxonomists to technically-skilled parataxonomists;
- Development of new sources of funding for supporting collections.
- Etc.

Taxonomy

For hundreds of years, primary species-occurrence data have been used for taxonomic and biogeographic studies. Data in museums and herbaria have primarily been used for the determination and description of new taxa. Collections were also used, however, for such things as studying pollination biology, evolutionary relationships, and phylogenetics. These uses continue, and with users now having access to data from a greater geographic range, they are able to expand on these studies

Taxonomic Research

There are thousands of published examples of uses of primary species-occurrence data in taxonomy and in the elucidation of new taxa and phylogenetic relationships. Species data in

museums are core to the study of basic taxonomy – the elucidation of new taxa and their descriptions. The world has about 1.4 million taxa already described (World Resources 1992) – nearly all based on collections in museums and herbaria. Many more still need to be described and thus one of the basic uses of species-occurrence data is the description and classification of plants, animals, algae, fungi, viruses, etc. Without these data, these processes could not continue.

Taxonomic projects are carried out at virtually every natural history museum and herbarium in the world with outputs in journals, monographs and electronically.

Examples:

- Biodiversity and Management and Utilization of West African Fishes is a project of ICLARM examining the taxonomy and phylogeny of fishes in Ghana and other West African states. <[+ x+1+](#)>;
- Cicadas of South-East Asia and the West Pacific – research from the Institute for Biodiversity Research and Ecosystem Dynamics of the Zoological Museum of Amsterdam (Duffels 2003). <<http://www.science.uva.nl/ZMA/entomology/CicadasSE.html>>
- The taxonomy of Vietnam's exploited seahorses (Syngnathidae) (Lourie, *et al.* 1999). <http://seahorse.fisheries.ubc.ca/pubs/Lourie_etal_vietnam.pdf>.
- HymAToL – a project aimed at constructing a large-scale phylogenetic analysis of the Hymenoptera of the world as part of the Tree of Life project. <<http://www.hymatol.org/about.html>>.
- Phylogeny. A project from the University of Alberta in Canada. <<http://www.deer.rr.ualberta.ca/library/phylogeny/Phylogeny.html>>.

Name and Taxonomic Indices

Primary species-occurrence data has been used to develop lists of names and taxa which are used in one way or another by most of the projects throughout this paper. In much the same way as dictionaries and thesauri are used in the spoken and written languages of the word, indexes of names and taxa are used for the language of biodiversity. Collections institutions use them as authority files for their databases, taxonomists use them to help determine the correct spelling and the place of original publication, and scientists and amateurs use them to find the correct spelling of a name of a species, its synonyms and other information. These indexes can vary from being just a list of names, to detailed lists that include taxonomic information, synonyms, place of publication, type specimen information, references to different uses of the names (taxonomic concepts), etc.

Examples:

- Species2000 <<http://www.species2000.org>>;
- Integrated Taxonomic Information System (ITIS) <<http://www.itis.usda.gov/>>;
- International Plant Name Index (IPNI) <<http://www.ipni.org/index.html>>;
- Electronic Catalogue of Names of Known Organisms (ECat) program of GBIF <<http://www.gbif.org/prog/ecat>>;
- Universal Biodiversity Indexer and Organizer (UBio) <<http://www.ubio.org/>>;
- Index Fungorum <<http://www.indexfungorum.org/>>;
- Index of Viruses <<http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/index.htm>>;
- Taxonomic Search Engine (TSE) <<http://darwin.zoology.gla.ac.uk/~rpage/portal/>>;
- Nomenclator Zoologicus <<http://ui0.mbl.edu/NomenclatorZoologicus/>>;
- Global Lepidoptera Names Index <<http://www.nhm.ac.uk/entomology/lepindex/>>;
- Tropicos <<http://mobot.mobot.org/W3T/Search/vast.html>>;

- Gray Card Index of Harvard University <<http://www.huh.harvard.edu/databases/>>.

Floras and Faunas

The publication of floras and faunas is one of the first outputs from the results of taxonomic research and their development is being greatly enhanced through access to species-occurrence data on-line. Most published floras and faunas include location information, and more often than not a simple mapped distribution. Traditionally, these maps were drawn by hand, and were invariably created without access to the totality of collections available. With distributed systems such as the GBIF Portal, and using a simple GIS, these maps can now be produced quickly and easily, and by having access to many more collections, are more likely to cover the totality of the distribution.

Examples:

- Flora of Australia online (ABRS, Canberra) <<http://www.deh.gov.au/biodiversity/abrs/online-resources/abif/flora/main/>>;
- Fauna of New Zealand (Manaaki Whenua Landcare Research) <<http://www.landcareresearch.co.nz/research/biodiversity/invertebratesprog/faunaofnz/>>;
- FaunaItalia <<http://faunaitalia.it/index.htm>> ;
- Phanerogamic Flora of the State of São Paulo (Brazil) <<http://www.cria.org.br/flora/>>.

Taxonomy and Ecological Biogeography

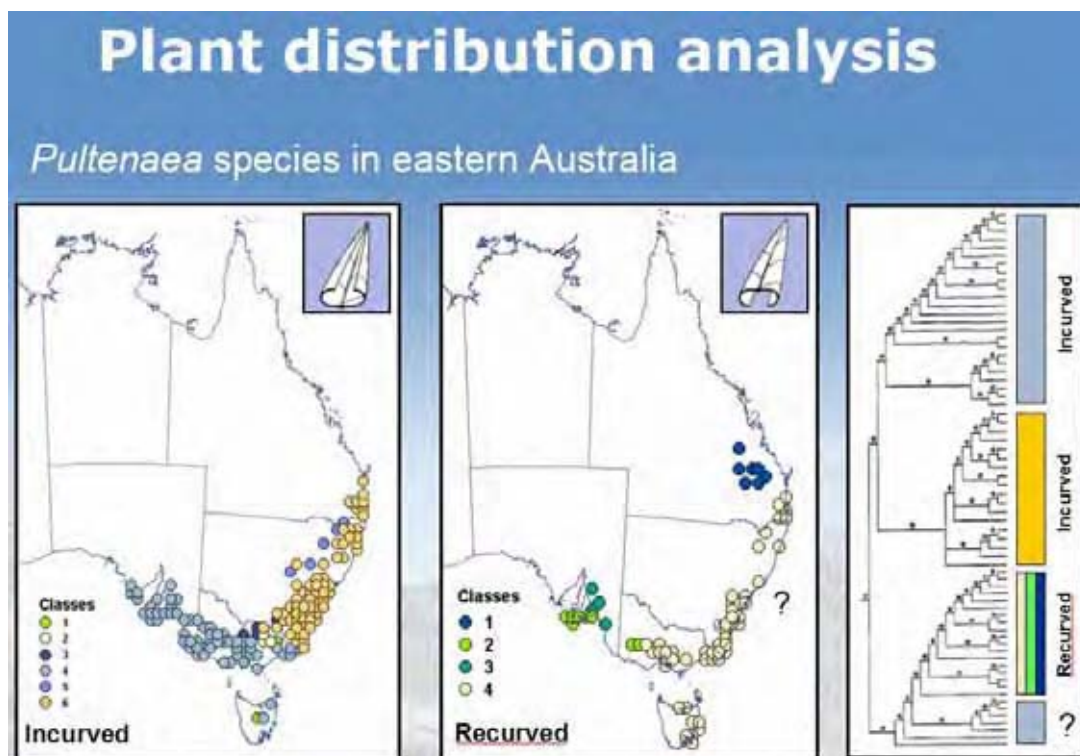


Fig. 1. Phylogenetic information from *Pultenaea* species in Australia showing geographic patterns related to leaf morphology. Phylogenetic groups were determined using cluster analysis from herbarium records with affinities hypothesized using leaf morphology and the phylogenetic cladogram derived from molecular data (right). Data

were collated through the Australian Virtual Herbarium (AVH) (CHAH 2002). Image from West and Whitbread (2004) with permission of the authors.

The availability of distributed data points from many collections agencies, now allows for quicker and more detailed studies, for example by looking at provenance differences, locations of collections with different characteristics (plotting location against leaf length for example), and the mapping of different taxonomic concepts. Many of the products mentioned below (Floras, Faunas, field guides, etc.) are the visible output from the basic taxonomic research.

Examples:

- A project at the Centre for Plant Biodiversity Research in Australia, maps patterns related to leaf morphology in phylogenetic groups of the genus *Pultenaea* (figure 1). Groups were identified on the basis of leaf morphology and a phylogenetic cladogram based on molecular data (Bickford *et al.* 2004, West and Whitbread 2004).
- Another project at the Centre for Plant Biodiversity Research, uses data obtained from 8 Australian herbaria accessed through the Australian Virtual Herbarium (CHAH 2002) to plot geographic patterns related to different taxonomic concepts (West and Whitbread 2004).

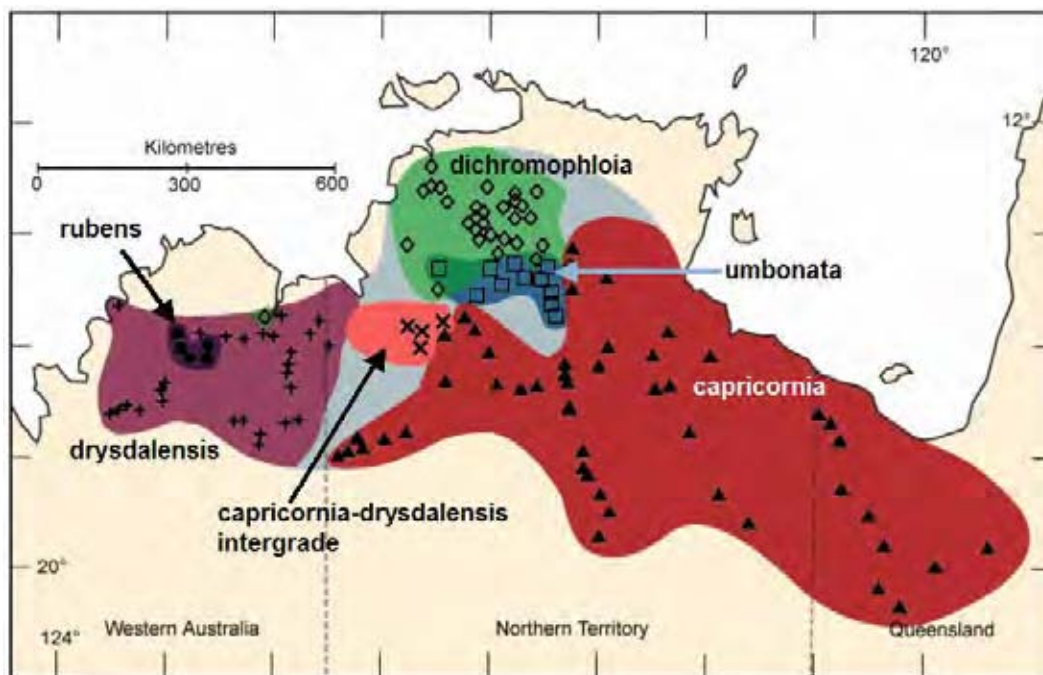


Fig. 2. Map showing different interpretations of a group of species in the genus *Corymbia* (previously part of *Eucalyptus*). Different taxonomic concepts of experts propose *C. umbonata* and *C. dichromophloia* encompassing the total distribution of the group as shown, as compared to another concept which interprets *C. dichromophloia* in a more narrow sense and recognises a number of other species as mapped here. Image from West and Whitbread (2004) with permission of the authors.

Field Guides

Most field guides incorporate a mapped distribution of the species under consideration. Again, like Floras and Faunas, they have traditionally included hand-drawn maps derived from the author's knowledge of the species. The availability of distributed species data now

makes the production of maps and the inclusion of distributional information, that much easier and far more accurate.

Examples:

- Birds of Argentina and Uruguay. A Field Guide (Narotsky and Yzurieta 2003).
- Dragonfly Recording Network
<<http://www.searchnbn.net/organisation/organisation.jsp?orgKey=6>>;
- Catalogue of the species of the Annelid Polychaetes of the Brazilian Coast (Amaral and Nallin 2004);
- Butterflies of North America
<<http://www.npwrc.usgs.gov/resource/distr/lepid/bflyusa/bflyusa.htm>>;
- Butterflies of Australia (Braby 2000);
- Tour 2 from GBIF Demonstration Project 2003: Access to multi-author rainforest inventories <<http://gbifdemo.utu.fi/>>.
- BumblebeeID – find British species by colour pattern.
<http://www.nhm.ac.uk/entomology/bombus/key_colour_british/ck_widespread.html>

Integrated electronic resources

The development of character-based databases, interactive keys, and digital imaging, along with the arrival of CD-ROMs and DVDs has led to the development of a number of integrated electronic resources.

Examples:

- PoliKey (an interactive key and information system for polychaete families and higher taxa) (Glasby and Fauchald 2003)
- Publications from the Expert Centre for Taxonomic Identification (ETI) produced using the Linnaeus II software (Shalk and Heijman 1996).
 - Searchable and Browsable Index to CD products produced using the Linnaeus software <<http://www.eti.uva.nl/Products/Search.html>>. Some examples include:
 - Catalogue of the Chalcicoidea of the World,
 - Birds of Europe,
 - Crabs of Japan,
 - Davalliaceae,
 - Fauna Malesiana, and
 - Fishes of the North-Eastern Atlantic and Mediterranean.
 - Arthropods of Economic Importance
 - Bats of the Indian Subcontinent
 - Key to Cotton Insects
- Publications using the Lucid Software (University of Queensland 2004):
 - Searchable Index to published products using the Lucid software. Searches can be conducted taxonomically, geographically and in a number of other ways <<http://www.lucidcentral.com/keys/keysearch.aspx>>. Examples include:
 - Key to Common Chilocorus species of India (J. Poorani). an economically important genus of lady beetles,
 - Key to the World Genera of Eulophidae Parasitoids (Hymenoptera) of Leafmining Agromyzidae (Diptera),
 - Key to Insect Orders,
 - Pest Thrips of the World.

- Publications using DELTA and IntKey (Dalwitz and Paine (1986).
 - Index to publications using DELTA and IntKey
 - <<http://biodiversity.bio.uno.edu/delta/www/data.htm>>. Some examples:,
 - Beetle – Elateroformia (Coleoptera) – families – (adults and larvae separate). Downloadable characters and descriptions for use in the Intkey program.
 - Braconidae (Hymenoptera) of the New World – subfamilies, genera and species > ,
 - Downloadable characters and descriptions for use in the Intkey program - in English and Spanish.
 - Commercial timbers (in English, German, French, and Spanish)
 - Polychaete families and higher taxa
- Publications using XID Authoring System
 - <<http://www.exetersoftware.com/cat/xid.html>>
 - Weeds of North America. A comprehensive weed identification reference for North America on CD, it contains 140 grass-like and 860 broadleaf weeds.
- CD-ROM Publications from the Australian Biological Resources Study (ABRS) and the Centre for Plant Biodiversity Research in Australia produced largely through use of Lucid Software (University of Queensland 2004)

Examples include

(<<http://www.deh.gov.au/biodiversity/abrs/publications/cds/index.html>>):

 - Acacias of Australia,
 - Mites in Soil,
 - AusGrass,
 - Spiders of Australia,
 - Australian Tropical Rainforest Trees and Shrubs
 - <http://www.anbg.gov.au/cpbr/cd-keys/rainforest-key/home_page.html>, and
 - Eucalypts of Southern Australia <<http://www.anbg.gov.au/cpbr/cd-keys/Euclid/>>.

Check lists and inventories

Species checklists for regions, national parks, etc. can now be produced almost automatically, and maintained through the use of distributed information systems. This is probably one of the least used, but most powerful use of a distributed system.

Examples:

- Checklist of Amphibian Species and Identification Guide. An online Guide for the Identification of Amphibians in North America north of Mexico.
 - <<http://www.npwrc.usgs.gov/narcam/idguide/>>;
- A Checklist of the Ants of Michigan
 - <<http://insects.ummz.lsa.umich.edu/fauna/MICHANTS.html>>;
- Checklist of the Amphibians and Reptiles of Rara Avis, Costa Rica
 - <<http://www.rara-avis.com/herplist.htm>>;
- Checklist and distribution of the liverworts and hornworts of sub-Saharan Africa, including the East African Islands <<http://www.oshea.demon.co.uk/tbr/tbrr3.htm>>;
- The Australian Mammal Audit (McKenzie and Burbidge 2002) was part of biodiversity audit for Australia

<http://audit.ea.gov.au/ANRA/vegetation/docs/national/FINAL_MAMMAL_REPORT.doc>.

Image Databases

The use of Image databases, especially of type specimens is reducing damage to natural history collections as taxonomists use images of the specimens, or of the labels, rather than borrowing specimens.

Examples:

- New York Botanical Garden Vascular Plant Type Catalog <<http://www.nybg.org/bsci/hcol/vasc/Acanthaceae.html>>;
- Parasite Image Library <http://www.dpd.cdc.gov/dpdx/HTML/Image_Library.htm>;
- Natural History Museum (London) Specimen label images <<http://atiniui.nhm.org/gallery/album33>>.

Phylogenies

The study of phylogenies, or evolutionary trees is enhanced by the use of primary species-occurrence data.

Examples:

- Tree of Life – a collaborative Internet project containing information about phylogeny and biodiversity <<http://tolweb.org/tree/phylogeny.html>>;
- The study of phylogenetic patterns in groups of *Pultenaea* (figure 1) (Bickford *et al.* 2004).

Parataxonomy

Parataxonomists are used in a number of developing countries to do preliminary sorting of collections. These parataxonomists rely on good species-occurrence data and products to be able to carry out their work efficiently and effectively.

Examples:

- Parataxonomists have been extensively used in the Guanacaste Conservation Area in Costa Rica (Janzen *et al.* 1993) <<http://www.unep-wcmc.org/forest/restoration/docs/CostaRica.pdf>>;
- Parataxonomists are being used to conduct biological surveys by the New Guinea Binatang Research Centre <<http://www.entu.cas.cz/png/parataxonomists.htm>>.

Automated Identification Tools

Automated identification tools that use pattern recognition followed by clustering, ordination or use of artificial neural network are being tested for use with insects, birds and frogs.

Examples:

- In Germany bees can be identified using pattern recognition with the Automatic Bee Identification Software (ABIS) <http://www.informatik.uni-bonn.de/projects/ABIS/ABIS_Contact.html>;
- In Japan, cicadas and grasshoppers are being identified using hand-held recorders to recognise calls using the Intelligent Bioacoustic Identification System (IBIS)

<http://www.elec.york.ac.uk/intsys/users/ijf101/research/acoustics/grasshoppers.shtml>>;

- In Britain, the Intelligent Bioacoustic Identification System (IBIS) is being used to identify bats
<http://www.elec.york.ac.uk/intsys/users/ijf101/research/acoustics/bats.shtml>>; as well as to identify sett occupancy in badgers underground
<http://www.elec.york.ac.uk/intsys/users/ijf101/research/acoustics/badgers.shtml>>;
- In Finland, sinusoidal modelling of birdcalls allows for the development of automated identification of birds (Härmä 2003).

Biogeographic Studies

Natural history collections contain a unique and irreplaceable record of the natural and cultural history of our world. Many of the specimens and ancillary data in collections were obtained prior to the major modifications of the landscape and they are irreplaceable (Chapman 1999, Page *et al.* 2004). Indeed, the collections are the fundamental database on the changing landscapes and patterns of species distributions (Page *et al.* 2004).

There are hundreds, if not thousands of biogeographic studies using species-occurrence data. Some use simple distributions within a grid, others link to environmental data layers such as climate and geology through environmental modelling tools, others look at various combinations to develop indices of diversity and endemism, relative abundance, etc. All such projects benefit from being able to access distributed data from multiple institutions. Examples will be included under individual headings below.

The use of environmental modelling software such as BIOCLIM (Nix 1986, Busby 1991) GARP (Stockwell and Peters 1999, Pereira 2002), and methods such as GLM (Austin 2002), GAM (Hastie and Tibshirani 1990), Decision Trees (Breiman 1984), and Artificial Neural Networks (Fitzgerald and Lees 1992), etc. to link individual locations of plants and animals to environmental criteria such as climate to produce maps of potential distribution have been around for more than 20 years. Because of the scale of environmental layers available at the time, some of the earlier studies looked at broad-scale distributions of groups of plants or animals, such as used with the Elapid Snakes (Longmore 1986), or more intensely on one species such as with *Nothofagus cunninghamii* (Busby 1984). Because of the nature of the software available at the time, and the paucity of good environmental layers, these studies were slow and took months to produce a model for just one species, and were often carried out at a scale that allowed for only broad conclusions to be drawn. The development of new software and vastly improved environmental layers (Hijmans *et al.* 2004) has meant that models can now be produced in limited time, allowing for more intensive studies of individual species, or studies on much larger numbers of species. Care, however, needs to be taken in using any of these modelling methods, and it is best to seek advice from experts before using them to ensure that the right model is being used for the right data etc. (Chapman *et al.* 2005).

Distribution Atlases

Traditional uses for geo-referenced primary species data have been for developing maps of species distributions and the development of distribution atlases. In the past, these have often been as a presence or absence within a geographic grid, from 5 km to 2.5-degree grids, or in a biogeographic region. Many of these have not been made available electronically.

Examples of mapping by grid or region include:

- Fife Bird Atlas (2 km grid squares) <http://www.the-soc.fsnet.co.uk/fife_bird_atlas.htm>;
- Atlas of the British Flora (Perring and Walters 1962) (10 km grid squares);
- Millenium Atlas of Butterflies in Britain and Ireland (Asher *et al.* 2001) (10 km grid squares);
- Ontario Herpetofaunal Summary Atlas (10 km grid squares) <<http://www.mnr.gov.on.ca/MNR/nhic/herps/about.html>>;
- The Introduction and spread of the Asian Long-horned Beetle in the north America is being studied using biogeographic analysis <<http://www.uvm.edu/albeetle/>> and Peterson *et al.* (2004) <http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PSH_AMN_2004.pdf>.
- Atlas of Australian Birds (1st edition) (Blakers *et al.* 1984) (10-minute grid squares);
- Atlas Florae Europaeae (50 km grid squares) <<http://www.biologie.uni-hamburg.de/b-online/ibc99/IDB/afe.html>>;
- Census of Australian Vascular Plants (Hnatiuk 1990) (97 biogeographic regions covering all of Australia);
- Moths of North America (Counties or States) <<http://www.npwrc.usgs.gov/resource/distr/lepid/moths/mothsusa.htm>>.

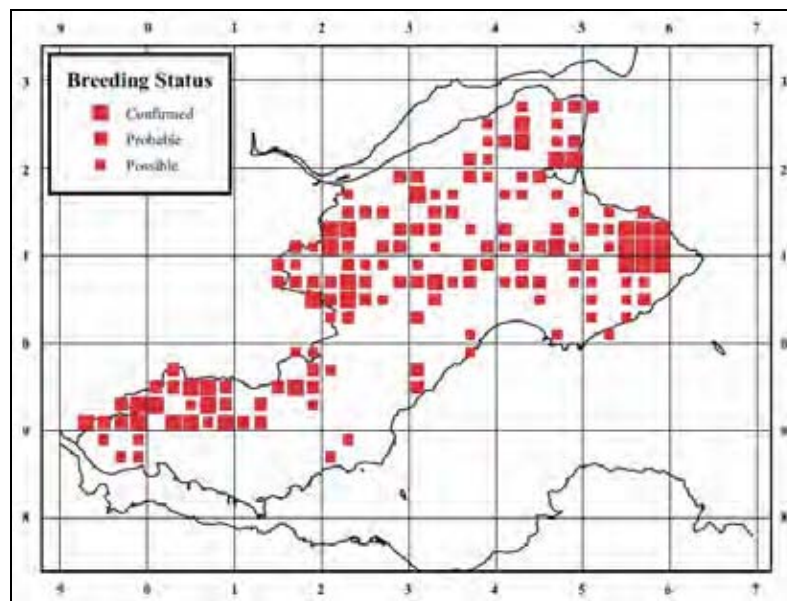


Fig. 3. Distribution of the Eurasian Curlew (*Numenius arquata*) in Fife, Scotland from the Fife Bird Atlas (Elkins *et al.* 2003) using 2 km grid squares. Map reproduced with permission of the authors.

Many of the early species distribution atlases were done by hand, and often without carrying out full geo-referencing. Mapping distributions in a grid could be carried out without a GIS and were easy to record merely as present or absent within each grid cell. The use of distributed database searches and Geographic Information Systems (GIS) now allows species distribution mapping and atlases to be produced much more accurately and with better presentation, and has allowed easier mapping of individual specimen records.

Examples of mapping individual records include:

- Atlas of Elapid Snakes of Australia (Longmore 1986);
- Protea Atlas Project (South Africa) <<http://protea.worldonline.co.za/default.htm>>;

- The New Atlas of Australian Birds <<http://www.birdsaustralia.com.au/atlas/>>;
- Tour 1 from GBIF Demonstration Project 2003: Reliability and consistency of Neotropical species distributions <<http://gbifdemo.utu.fi/>>;
- Atlas of the Birds of Mexico (Navarro *et al.* 2003).

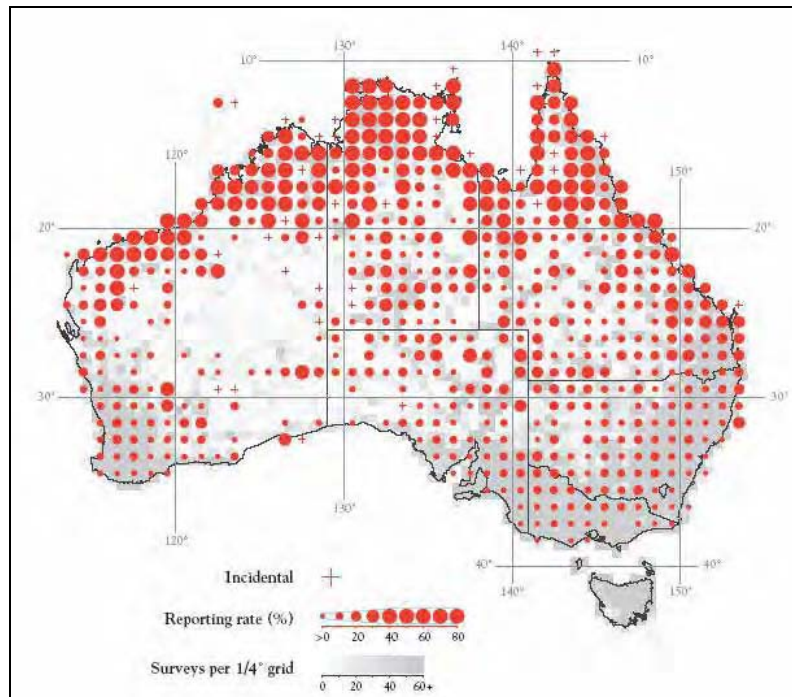


Fig. 4. *Distribution of the Rainbow Bee-eater from The New Atlas of Australian Birds (Barrett et al. 2003). Records are recorded as point records and mapped as a summary in 1-degree grid squares (red) and on 0.25-degree grid squares (grey).*

Species Distribution Modelling

In the mid 1980s, the concept of environmental species distribution modelling using environmental data such as climate, started to become possible with the development of computer software such as BIOCLIM (Nix 1986, Busby 1991). Since then, many new modelling methodologies and programs have been developed, including Generalised Linear Models (GLM) (Austin 2002), Generalised Additive Models (GAM) (Hastie and Tibshirani 1990), Genetic Algorithm for Rule-set Production (GARP) (Stockwell and Peters 1999, Pereira 2002), DOMAIN (Carpenter *et al.* 1993) and many, many others. These programs were stand-alone programs, but the availability of World Wide Web in 1994, saw the development of modelling on the Internet – firstly with BIOCLIM and GARP (Boston and Stockwell 1995), and later with modifications of these and other programs.

The development of these modelling techniques opened up primary species-occurrence data to many more uses. One of the main drawbacks of these data are their lack of comprehensiveness and completeness, and the use of models allows for gaps in the distributional knowledge of species to be filled. There are now many projects using modelling techniques for determining the potential distributions of species under present-day climatic conditions given various constraints, under altered climatic conditions following climate change, and under past climatic conditions in earlier epochs. Some of these uses will be covered under more specific topics below, but

Examples:

- Atlas of Elapid Snakes of Australia (BIOCLIM) (Longmore 1986);
- Atlas of Vertebrates Endemic to Australia's Wet Tropics (BIOCLIM) (Nix and Switzer 1991);
- Use of Environmental Gradients in Vegetation and Fauna Modelling (GLM) (Austin 2002);
- Potential distribution of *Anoplophora glabripennis* (Asian Long-horned Beetle) in North America (GARP) (Peterson *et al.* 2004);
- Predicting distributions of Mexican birds (GARP) (Peterson *et al.* 2002b);
- In Africa, tsetse fly habitats were modelled using species data and remotely-sensed vegetation data (Robinson *et al.* 1997).

Atlas of Elapid Snakes of Australia

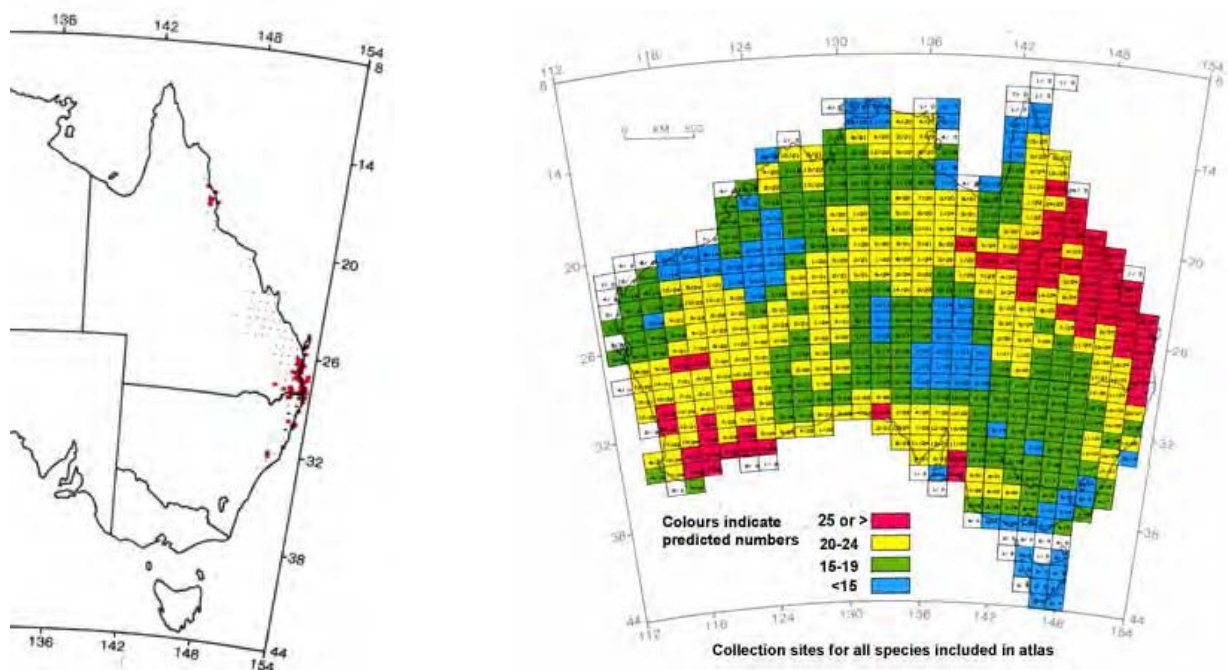


Fig. 5. Left-hand image - Potential distribution for *Tropidechis carinatis* in Australia. Red stars indicate known collections, dots show modelled distribution. Right-hand image shows predicted numbers of species in each 1° x 1.5° cell. From Longmore (1986) with permission of Australian Biological Resources Study.

The Atlas of Elapid Snakes (Longmore 1986) was a result of a pilot project conducted with the Australian Museum in 1982 to examine uses for geo-referenced primary species data. In 1983, the Australian Bureau of Flora and Fauna (now the Australian Biological Resources Study), decided that it was wasting resources by funding the collection of new species records without first utilising data already held by museums. Data for 17,000 records were then collected from all the major Australian museums, integrated and modelled using the bioclimatic modelling software, BIOCLIM (Nix 1986, Busby 1991). Many of the data were in a poor state of curation and required extensive data validation and cleaning prior to use. The Atlas contained maps for all 77 species of front-fanged, venomous terrestrial snakes (the family Elapidae) in Australia and was one of the first attempts to collate, geo-reference, and document all records of an animal group for purposes of biogeographic study. The project also saw the first detailed publication of the software program, BIOCLIM (Nix 1986).

Environmental data layers for use in bioclimatic modelling were still quite primitive. Twelve climate parameters were used at a scale of 0.5-degree resolution. Species data were geo-referenced as accurately as possible, and altitude determined to the nearest 50 m. The species were modelled using the 5-95 and 100 percentile ranges and mapped at a continental scale (figure 5).

Predicting new species distributions

By using species-occurrence data in conjunction with species modelling tools, it is possible for additional locations of species to be identified. In other cases, species modelling has identified disjunctions in climate profiles that have indicated that two species are present where only one was previously known.

Examples:

- Museum collections as well as new survey data were used to predict reptile diversity in Madagascar and were successful in predicting locations of new chameleon species (Raxworthy *et al.* 2003);
- In Australia, new locations of a rare *Leptospermum* species (Myrtaceae) were identified using species modelling (Lyne 1993)
<<http://www.anbg.gov.au/projects/leptospermum/leptospermum-namadgiensis.html>>.

Studying species decline

By using locality information and collection information such as date of collection, primary species data can help in the understanding of species declines over time.

Examples:

- AmphibiaWeb (Wake 2004) <<http://amphibiaweb.org/>>; Species Decline: Contaminants as a Contributing Factor. Patuxent Wildlife Research Center Database <<http://www.pwrc.usgs.gov/pattee/select.htm>>;
- The Red List Index has developed a tool for measuring global trends in the status of biodiversity (Butchart *et al.* 2004).
<http://www.birdlife.org/print.html?url=%2Fnews%2Fpr%2F2004%2F10%2Fred_list_indices.html>;
- Australian Terrestrial Biodiversity Assessment is part of the Australian Natural Resources Atlas v. 2.0
<http://audit.ea.gov.au/ANRA/vegetation/vegetation_frame.cfm?region_type=AUS®ion_code=AUS&info=bio_asses>.

Species Diversity and Populations

The study of species diversity, species density and richness is a discipline that is being aided enormously by the increasing availability of species-occurrence data. In the past, these types of studies required months, if not years of data collection and preparation, and usually concentrated on the data available from just a few museums or herbaria and thus seldom covered the totality of the data. This new availability of data through distributed systems has meant that new tools are being developed to cater for the increases in data availability and to allow for more rapid analysis and assessment. As a result, the data can be used more

effectively in biodiversity assessment projects, in conservation assessment and for regional planning and management.

The increased availability of data is allowing for improved modelling and distribution of associations and populations leading to improved understandings of species and how they interact with their environments. This is allowing for better management of populations, and understandings of threatened species and communities. This improved understanding, for example, is now allowing Australia to list threatened ecological communities as well as species (DEH 2000, 2004).

Species Diversity, Richness and Density

The study of species richness, density and abundance and the identification of centres of endemism have been key areas of research in biodiversity over the past 20 years. More recently, they have been integrated into conservation assessment and planning and species protection. In many cases, species diversity, and richness are used as surrogates for measuring biodiversity.

Species Richness Tools

New tools are being developed to assist in assessment of species richness and endemism and for use as planning tools for conservation assessment.

Examples:

- WorldMap uses species distribution data to produce species richness maps, which can then be used to carry out further analyses. (Williams *et al.* 1996) <<http://www.nhm.ac.uk/science/projects/worldmap/index.html>>;
- Australian Heritage Assessment Tool, under development at the Australian Department of the Environment and Heritage, can quickly generate maps of richness and endemism for a broad range of Australian plant, vertebrate and invertebrate taxa through an easy to use interface (figure 4);
- Pattern Analysis tools such as PATN (Belbin 1994) can be used to identify patterns in species diversity and endemism <<http://www.patn.com.au/>>;
- EstimateS is another software package for estimating species richness. (Colwell 2000) <<http://viceroy.eeb.uconn.edu/estimates>>;
- Species Richness bibliography <<http://www.okstate.edu/artsci/botany/ecology/richness.htm>>.

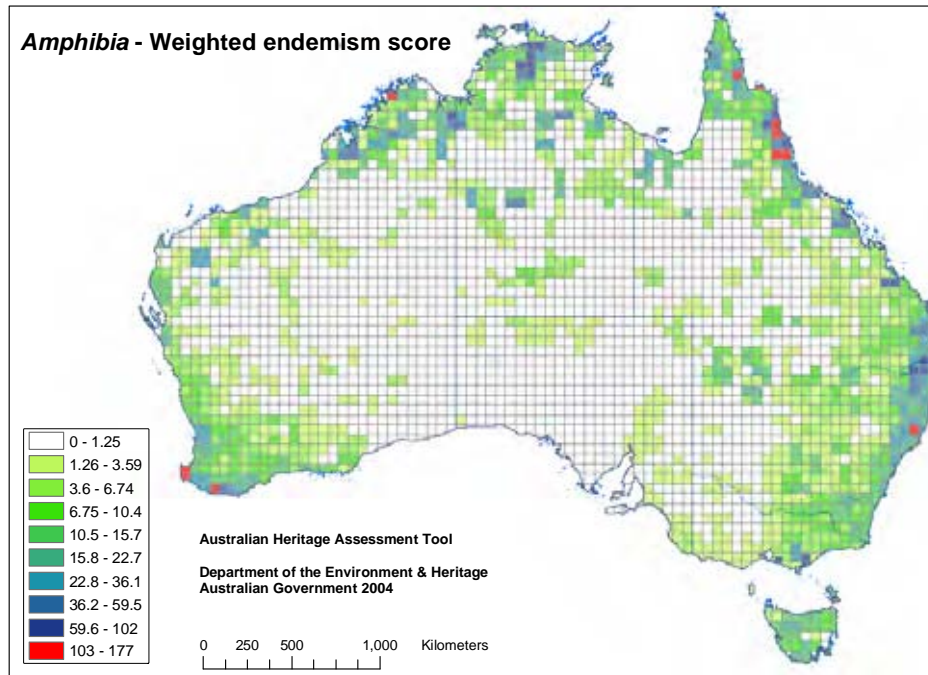


Fig. 6. Endemism in Australian frogs showing peak areas for frog endemism highlighted in red. Image from the Australian Heritage Assessment Tool; published with permission of Cameron Slatyer and Dan Rosauer, Australian Department of the Environment and Heritage, 2004.

Biodiversity Hotspots

Biodiversity hotspots or centres of endemism are regarded as the world's biologically richest and most important areas for conservation (Mittermeier *et al.* 2000). Conservation International has been conducting a program to assess those areas of the world regarded as the most "species rich".

Examples:

- Conservation International identifies the 25 most threatened biodiversity rich areas of the world (Myers *et al.* 2000)
<<http://www.biodiversityhotspots.org/xp/Hotspots>>;
- Birdlife International's Endemic Bird Areas of the world (Stattersfield *et al.* 1998)
<http://www.birdlife.net/action/science/endemic_bird_areas/>;
- Biodiversity hotspots in Australia
<<http://www.deh.gov.au/biodiversity/hotspots/index.html>>;
- The Millenium Atlas of Butterflies is mapping the species richness of butterflies across the United Kingdom
<<http://www.butterfly-conservation.org/index.html?/bnm/atlas/index.html>>.

Patterns of Species Richness

Species richness studies are conducted from the size of one vegetation community to a global scale. Most species richness studies have implications for conservation, the identification of hot spots as mentioned above and the identification of priority areas for conservation.

Examples:

- A study in central Brazil looked at the richness and abundance of caterpillars of one genus of plant in the cerrado (savannah-like) vegetation (Andrade *et al.* 1999)

http://www.scielo.sa.cr/scielo.php?pid=S0034-77441999000400005&script=sci_arttext&tlng=en>;

- A study in Africa is looking at species richness and endemism of insects in sub-Saharan Africa (Miller and Rogo 2001);
- Species richness and endemism in South American bird species was used to plan a network of reserves (Fjeldsa and Rahbek 1997);
- The geographic relationships and constraints on species richness were studied using mid-domain effects (Colwell and Lees 2000);
- Examining spatial patterns at the community level (Ferrier *et al.* 2002).

Studying Individual species

Species richness studies of single species – knowing where it occurs, and where it moves, and the densities of individual populations, can aid in the conservation of that species. By using historical data, changes in patterns of movement can be examined.

Examples

- The density of Elephants in the forests of central Africa is being studied using Geographic Information Systems (Michelmore 1994).

Evolutionary patterns

One of the aspects of species richness studies is the detection of patterns of endemism and richness. By looking at the patterns of species concentrations and endemism, historical evolutionary patterns can be determined.

Examples:

- In a study of conservation in Africa, Brooks (2001) examined four groups of animals – mammals, birds, snakes and amphibians and modelled species richness against environmental conditions such as primary productivity potential evapotranspiration, solar radiation, temperature, and rainfall.

Population Modelling — Population Viability Analysis

The modelling of populations can help track the dynamics of the population, and assist in determining a minimum area for conservation, and examine interactions with predators and prey, etc. Species observational data and data from intensive survey is an essential tool for these studies. Population Viability Analysis (PVA) was originally used to determine how large a population must be to have a reasonable chance of survival for a reasonable length of time.

Examples:

- At the Centre for Resource and Environmental Studies in Canberra, detailed studies have been conducted on populations of a small threatened marsupial – Leadbeater's Possum (*Gymnobelideus leadbeateri*) in the forests of northern Victoria. (Lindenmeyer and Possingham 1995, 2001. Lindenmeyer and Taylor 2001) <http://incres.anu.edu.au/possum/possum.html>>;
- Applied Biomathematics® is using the RAMAS software package to model extinction risk in birds through use of Population Viability Analysis <http://www.ramas.com/birds.htm>>;
- Many studies in China have used Population Viability Analysis to examine minimum reserve size for maintenance of viable populations of the Giant Panda (*Ailuropoda melanoleuca*) (Zhou and Pan 1997);

- An annual census of Southern Elephant Seals is conducted on sub-Antarctic Macquarie Island on the 15th October every year, and annual populations' estimates made (Burton 2001). It is estimated that around one-seventh of the world's populations of Elephant Seals live on the island, and that they forage over vast areas of the southern ocean from Heard Island in the west to the Ross Sea in the east <<http://www.aad.gov.au/default.asp?casid=3802>>.

Species Inter-relations

The study of species interactions is another area where species-occurrence data is essential. Such inter-relationships can include parasitic relationships, symbiotic relationships between species of animals, species of plants or between animals and plants; predator-prey relationships, competition, etc.

Examples:

- A project is being conducted in the Guanacaste Conservation Area in Costa Rica to make an inventory of Eukaryotic parasites of vertebrates (Brooks 2002) <http://brooksweb.zoo.utoronto.ca/FMPro?-DB=CONTENT.fp5&-Format=intro.html&-Lay=Layout_1&-Error=err.html&content_id=1&-Find>;
- A project at Madang, in Papua New Guinea looked at host specificity of insect herbivores on 60 species of rainforest trees. The project needed to cross-reference data on habitats, hosts, insect species, patterns of host use, and sampling events (Basset *et al.* 2000);
- The Parasite Database at the University of Toronto maintains information on parasite-host relationships <<http://brooksweb.zoo.utoronto.ca/index.html>>;
- A project at the European Network for Biodiversity Information (ENBI) in collaboration with African countries, is studying Afrotropical Ceratitidine Fruit Flies using a queryable web site on species distribution of insects and host plants. <<http://projects.bebif.be/enbi/fruitfly/>>
- Another study in Costa Rica is looking at the parasites of freshwater turtles (Platt 2000) <<http://brooksweb.zoo.utoronto.ca/pdf/Neopolystoma%20fentoni.pdf>>;
- In Canada, the predator-prey relationship between the nemertean (*Crebatulus lacteus*) and the soft-shell clam (*Mya arenaria*) is being studied (Bourque *et al.* 2002) <http://pubs.nrc-cnrc.gc.ca/cgi-bin/rp/rp2_abst_e?ciz_z02-095_80_ns_nf_cjz>;
- The World Federation of Culture Collections (WFCC) is supplying data via GBIF on interactions between parasites and hosts for many species <<http://wcdm.nig.ac.jp/hpcc.html>> as is the Belgian Co-ordinated Collections of Micro-organisms (BCCM) <<http://wcdm.nig.ac.jp/hpcc.html>>.

Protecting Communities

In Australia, new environmental protection legislation (DEH 2000) now allows for the listing of threatened communities in a similar way to the listing of threatened species. Communities can be listed as: critically endangered, conservation dependant or extinct in the wild and there are severe penalties for any significant impact on them. Primary species-occurrence data are used to determine boundaries and definitions prior to listing (Chapman *et al.* 2001).

Examples:

- Riverine aquatic protected areas: protecting species, communities or ecosystem processes? (Koehn 2003).

Life Histories and Phenologies

The study of life histories of both plants and animals is benefiting from the availability of species-occurrence data. The use of primary species data also aids the study of phenologies – being able to relate collections and records to the date and time of occurrence.

Life History Studies

Museum collections are a logical resource for life history studies. As stated by Pettitt in 1991

“Using existing collections for such studies often enables large amounts of data to be accumulated in a short time on such things as fecundity/mortality patterns, host-parasite relationships, estimates of breeding seasons, micro-growth increments (many organisms show growth layers when sectioned, such as the 'rings' of a tree, and these can be used to study past environmental conditions), food pests, life-cycle duration, larval growth pattern, migration (museum collections have been used to locate locust outbreak sites and to track traditional migration patterns), species that mimic other animals, and other polymorphisms, plant fecundity, flowering and fruiting dates, periods of dormancy, and correlations of plant growing sites with rainfall or altitude.” (Pettitt 1991).

Many animals and plants have completely different life stages, and species-occurrence data can supply a wealth of information on the relationship between different stages in the life cycle, and geographic locations or times of the year.

Examples:

- In the study of the North American Wood Stork (*Mycteria americana*) in Florida, museum collections were used to show that clutch sizes had not significantly declined since 1875 (Rogers 1990). Herons and egrets have also been studied <http://web8.si.edu/sms/irlspec/Cl_Aves3.htm>;
- Wingpad development in Plecoptera was studied in Italy using museum collections (Zwick 2003) <<http://www.unipg.it/maystone/PDF%202001%20proc/ZWICK2%20IJM%20proceedings.pdf>>.

Phenology

Phenology is the study of the timing of naturally occurring events and their relationship of biotic and abiotic variables. Examples include the flowering of plants, arrival and departure times of birds, the outbreak of plagues of locusts, the time of egg laying by monotremes and birds, etc. Primary species data are a major resource of information that can be used in phenological studies.

Examples:

- The study of the time of egg-laying of the codling moth (*Cydia pomonella*) an important pest of apples and pears, is important in determining times of spraying, etc. <http://www.ipm.ucdavis.edu/PHENOLOGY/ma-codling_moth.html>;
- In Kansas, a database of the times of flowering of wildflowers and grasses has been compiled <<http://www.lib.ksu.edu/wildflower/season.html>>;
- In the United States, the flight speed and rate of migration of birds is being studied <<http://www.npwrc.usgs.gov/resource/othrdata/migratio/speed.htm>>;

- Species data are being used in phenological studies of turtle nesting and migration <<http://www.natureserve.org/explorer/servlet/NatureServe?searchName=Chelonia+mydas>>.

Endangered, Migratory and Invasive Species

Endangered, migratory and invasive species are three groups of species regarded as key groups in biodiversity management. Indeed, in Australia, they are legislated as “nationally significant” (DEH 2000). Species-occurrence data are essential for the understanding and management of these groups of species in the environment.

Endangered Species

Endangered species provide many challenges to biogeographers, modellers and conservation biologists. There are usually so few records that environmental modelling techniques seldom work well. However, threatened species are essential components of any conservation program and species-occurrence records often provide the only available information. Primary species-occurrence databases are important for the identification of endangered species, identifying the reasons why they are endangered, for identifying external factors affecting the species, and for assisting in the development of species recovery plans.

Examples:

- IUCN Red List of Threatened Species <http://www.redlist.org/>.
- Endangered Species Program of the U.S.A. – U.S. Fish and Wildlife Service <<http://endangered.fws.gov/>>;
- Threatened Species Program – Australian Department of the Environment and Heritage <<http://www.deh.gov.au/biodiversity/threatened/index.html>>.

Species Recovery Plans

Species Recovery Plans are becoming an integral part of threatened species management in many countries.

Examples:

- Recovery Plan for the Angle-stemmed Myrtle (*Austromyrtus gonoclada*) Queensland Parks and Wildlife Service <<http://www.deh.gov.au/biodiversity/threatened/publications/recovery/a-gonoclada/index.html>>
- Threatened Species Recovery Plans Australian Department of the Environment and Heritage <<http://www.deh.gov.au/biodiversity/threatened/recovery/list-common.html>>;
- Threatened Species Recovery Plans New Zealand Department of Conservation <<http://www.doc.govt.nz/Publications/004~Science-and-Research/Biodiversity-Recovery-Unit/Recovery-plans.asp>> ;
- Recovery Plan Summaries from Environment Canada <http://www.speciesatrisk.gc.ca/publications/plans/default_e.cfm>.

Threats

The study of threats to endangered species can also be enhanced through the use of primary species data – especially when those threats are other species such as predators or competitors. In Australia, key threatening processes are listed under legislation, and include

such things as feral goats, the root-rot fungus (*Phyophthora cinnamomi*), the Fire Ant (*Solenopsis invicta*), etc.

Examples:

- Threat Abatement Plans – Australian Department of the Environment and Heritage <<http://www.deh.gov.au/biodiversity/threatened/tap/index.html>> ;
- Threats to Albatrosses and Giant-petrels <<http://www.deh.gov.au/biodiversity/threatened/publications/recovery/albatross/index.html>>;
- The introduction of the red imported fire ant, *Solenopsis invicta*, has caused a reduction in biodiversity of Australian native flora and fauna <<http://www.deh.gov.au/biodiversity/threatened/ktp/fireant.html>>.

Species Decline

The study of the decline in species numbers and distributions is an important step in preventing future endangerment and extinction in species and species habitats. Species-occurrence databases are an important information source for the study of both past declines and for monitoring current species numbers for prevention of future declines.

Examples:

- AmphibiaWeb <<http://www.amphibiaweb.org/declines/declines.html>>;
- FrogLog – Newsletter of the Declining Amphibian Populations Taskforce <<http://www.open.ac.uk/daptf/froglog/>>;
- At Cornell University in the United States, the status of birds are monitored to identify declining species <<http://www.sesc.k12.ar.us/2000backeast/ENatHist/Members/BryanM/page%202.htm>>;
- Predicting risk of extinction in declining species (Purvis *et al.* 2000) <<http://www.bio.ic.ac.uk/evolve/docs/pdfs/Purvis%202000%20PRSLB.PDF>>.

Invasive species and translocation studies

The spread of invasive alien and translocated species is one of the biggest environmental problems faced by most countries today. It is regarded by the Convention on Biological Diversity as the second most important threat to biodiversity after habitat change (CBD 2004). It is estimated that there are as many as 120,000 introduced species in the six countries made up of the United States, United Kingdom, Australia, India, South Africa and Brazil, alone (Pimentel 2002). Of these, perhaps 20-30% is now regarded as a pest species. The cost in economic loss of the 30,000 non-indigenous species in the United States has been estimated at close to \$123 billion a year (Pimentel *et al.* 1999, 2000).

Not all introduced species become invasive. In the history of the United States it is estimated that approximately 50,000 non-indigenous species have been introduced (Pimentel *et al.* 1999). Many of these of these have been used as food crops, livestock and farmed animals such as cattle and poultry, pets, biological control agents, landscape restoration, etc. However, those that have become pests cost the world a lot of resources every year in lost production, control and disease.

Preventing future invasions and predicting the impact of already introduced species requires accurate identifications and information on the natural distributions and ecological requirements of those species as well as associated species that may have positive or negative impacts with them (Page *et al.* 2004). The availability of species-occurrence data from

different countries through projects such as GBIF, allows researchers to identify the native locations of invasive species, determine the niche characteristics in the form of climatic and environmental requirements, and then use this information to predict likely spread in the country of introduction.

It also allows researchers to look at the distribution of possible biological control species, and to use this information to examine the possible spread and environmental limitations of these before introduction.

The availability of this information now makes possible, studies into invasive species and biological control agents that has not been previously possible, and this alone more than justifies the costs of projects such as GBIF.

There are many studies already using such information, and links to over 80 case studies can be seen on the web site of the Convention on Biological Diversity at <http://www.biodiv.org/programmes/cross-cutting/alien/cs.aspx>.

Example:

- The Global Invasive Species Program (GISP) in an On-line toolkit that “provides advice, references, and contacts to aid in preventing invasions by harmful species and eradicating or managing those invaders that establish populations” <http://www.cabi-bioscience.ch/wwwgisp/gtsum.htm>;
- Predicting the Geography of Species Invasions using Ecological Niche Modeling (Peterson 2003) <http://www.specifysoftware.org/Informatics/bios/biostownpeterson/P_ORB_2003.pdf>;
- In Kenya, the process of weed invasions have been tracked using herbarium specimens, showing that the regional spread of weeds in Kenya was correlated with the change in agricultural systems (Stadler *et al.* 1998);
- The spread of invasive Argentine Ants (*Linepithema humile*) across the United States over the past 100 years was studied by Suarez and others (2001) using both museum collections, and observations <http://www-biology.ucsd.edu/news/article_051500.html>;
- In New Zealand, bioclimatic prediction is being used to monitor the potential distribution of weeds prohibited entry to New Zealand (Panetta and Mitchell 1991);
- In North America studies are being carried out on the introduced Saltcedar (*Tamarix ramossisima*), which is becoming a major pest species in arid areas of Mexico where it is a huge user of water, aggressively replaces native riparian vegetation, and reduces habitat for birds and other animals. Distributions are being modelled in native and introduced habitats to assist planning in control and eradication (Soberón 2004);
- In Brazil and North America, the invasive potential of *Homalodisca coagulata* an insect vector of a bacteria of orchard-based crops was studied using distribution models with GARP (Peterson *et al.* 2003a);
- In Australia, invasive species are now listed under legislation and species-occurrence data are used to track their spread and to monitor their control <<http://www.deh.gov.au/biodiversity/invasive/index.html>>;
- Species distribution models were used to assess the invasive risk of several bird and insect species (Peterson and Vieglais 2001) <http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PV_B_2001.pdf>;

- Timely identification of pests can reduce need for costly control programs <http://www.bionet-intl.org/case_studies/case17.htm>.
- Harlequin Ladybird (*Harmonia axyridis*) study – a survey of an invasive species in the UK <www.harlequin-survey.org>

Arthropods and Annelids.

Approximately 4,500 arthropod species (2,582 species in Hawaii and more than 2,000 in the continental United States) have been introduced to the United States (Pimental 1999). In addition many aquatic invertebrates and earthworms have arrived. According to Pimental loc. cit., about 95% of the introductions were accidental.

Examples:

- North American Non-Indigenous Arthropod Database (NANIAD) is an on-line database of over 2,000 species of non-indigenous arthropods introduced into the United States of America <<http://www.invasivespecies.org/NANIAD.html>>.

Ballast Water

Ballast water in ships is a major source of introduced alien species into coastal habitats around the world. The identification of these species is an international problem as they may arise from anywhere in the world. The ability to use on-line primary species databases provides a major step forward in the identification and eventual regulation and control of these species.

Examples:

- The Northern Pacific seastar (*Asterias amurensis*) has virtually wiped out a species of shellfish and is a major threat to the marine environment. It is also adversely affecting the Tasmanian and Western Australian fisheries. It was not identified as an introduced species until 1992, and thus attempts to control it were delayed. Distributed primary species databases may help to prevent such delays occurring in the future <<http://www.fish.wa.gov.au/hab/broc/invasivespecies/seastar/>>;
- The Zebra Mussel (*Dreissena polymorpha*) originated in Poland and in the former Soviet Union, and after introduction in Ballast water are now causing problems throughout northern Europe and the United States, including in the Great Lakes between Canada and the United States <<http://nas.er.usgs.gov/taxgroup/mollusks/zebramussel/>>;
- In Australia, the Ballast Water Management Strategy uses species-occurrence data to identify, for example, where ballast water should not be taken on because of 'hot spots' of particular species that may become pests <<http://www.affa.gov.au/content/output.cfm?ObjectID=6F3A6281-9705-4878-9FA6836B5D6D5814>>.

Biological control of pests

The use of biological control agents to control pests has been in operation for around 50 years, and their use is increasing. Species-occurrence data are used to help find suitable biocontrol agents and to monitor their effectiveness and possible spread.

Examples:

- Biocontrol of mealybugs in South Africa <http://www.bionet-intl.org/case_studies/case2.htm>;
- Taxonomy is used in the selection of bio control agents in Hawaii <http://www.bionet-intl.org/case_studies/case15.htm>;

- Weevils are being used to control *Eichhornia crassipes* in Australia and elsewhere <<http://aquat1.ifas.ufl.edu/hyacin.html>>;
- Rabbits are controlled in Australia using various virus species <<http://www.csiro.au/communication/rabbits/qa1.htm>>.

Biological control gone wrong

The use of biological control agents must be controlled, otherwise disasters can occur. Species-occurrence data can be used to study locations of possible biological control agents, and to predict their possible spread in the proposed country of introduction. Not all biological control introductions in the past have worked.

Examples:

- In Australia, the Giant Cane Toad (*Bufo marinus*) was introduced into Australia in 1935 to control two introduced pests of the sugar cane industry – the Grey-backed cane beetle and the Frenchie beetle. CSIRO in Australia is mapping the spread through museum records and observations <<http://www.csiro.au/index.asp?type=faq&id=CaneToadControl&stylesheet=sectorInformationSheet>>;
- Many species have been introduced into Australia and South Africa to control *Lantana* species. The majority of these have not worked for a number of reasons, although some have worked in Hawaii and elsewhere. Different biological control agents have different effects on the different phenotypes of *Lantana* occurring in Australia, and the use of species-occurrence data to map the origins and spread of those phenotypes and the relationships of the bio-control agents in those areas can help improve success rates (Day and Nesser 2000).

Opuntia species in Mexico and the biological control agent *Cactoblastis cactorum*

Opuntia is one of the most used genera of plants in Mexico and Central America (Soberón *et al.* 2001), and is 10th in agricultural importance in Mexico (Soberón *et al.* 2000). The moth *Cactoblastis cactorum* is one of the best-known examples of a successful biological program when it was used in Australia in the control of *Opuntia* species in Queensland and northern New South Wales (Debach 1974). Fears have now arisen about the introduction of the *Cactoblastis* moth into Mexico, and the Commission on the Conservation and Use of Biodiversity in Mexico (Conabio) is modelling the potential spread and impacts of the moth there.

Examples:

- Using species-occurrence data and species distribution modelling to examine the potential spread and impact of *Cactoblastis cactorum* on the more than 90 species of native cactus species in Mexico and North America (Soberón *et al.* 2001) <<http://www.fcla.edu/FlaEnt/fe84p486.pdf>>.

Studying coevolutionary patterns

Museum collections have even been used to examine the rapid evolutionary response and adaptation of weeds to new environments.

Examples:

- In North America, studies on the co-evolution of parsnip (*Pastinaca sativa*) with the parsnip web worm (*Depressaria pastinacella*) have examined seeds from herbarium specimens to compare chemical co-evolution of the plants with the insect

(Berenbaum and Zangerl 1998).

<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24890> .

Migratory Species

Migratory species, virtually by definition, range across political boundaries and thus their study requires data from a range of jurisdictions. In the past, it has been difficult to obtain data from areas of a species range that one may be studying from outside the researcher's own country. The availability of distributed data systems is now allowing for new opportunities for migratory species studies. Various agreements are now in place around the world to track and monitor migratory species and to exchange information, including species-occurrence data.

Examples:

- Convention on Migratory Species (Bonn Convention) <<http://www.cms.int/>>;
- Japan-Australia Migratory Bird Agreement (JAMBA) <<http://www.austlii.edu.au/au/other/dfat/treaties/1981/6.html>> and China-Australia Migratory Bird Agreement <<http://www.austlii.edu.au/au/other/dfat/treaties/1988/22.html>>;
- African-Eurasian Migratory Water Bird Agreement <<http://www.unep-wcmc.org/AEWA/index2.html>>;
- Global Register of Migratory Species (GROMS) <<http://www.groms.de/>>.
- Migratory Birds know no Boundaries An extensive information resource from Israel on migratory birds <<http://www.birds.org.il/>>

Tracking Migratory Species

The tracking of migratory species and where they move has been an ongoing process for many years. One of the problems in the past has been the lack of access to species-occurrence data. With the new availability of species-occurrence data, data from all the range states can be combined to track and monitor changes in patterns of behaviour, decline in numbers, life spans etc. Tracking may be through observation and counts, through banding and recapture, through use of satellite tracking devices, or by use of radioactive isotopes.

Examples:

- European Union for Bird Ringing <<http://www.euring.org/>>;
- Australian Bird and Bat Banding Scheme (ABBBS) <<http://www.deh.gov.au/biodiversity/science/abbbs/>>;
- The Monarch Butterfly (*Danaus plexippus*) migration is tracked from Mexico to the United States each year through the use of banding <http://www.uen.org/utahlink/activities/view_activity.cgi?activity_id=2030>;
- Hydrogen isotopes (heavy water or deuterium) are being used to track Monarch Butterfly (*Danaus plexippus*) breeding and feeding grounds (Wassenaar and Hobson 1998) <<http://whyfiles.org/083isotope/2.html>>;
- In Malaysia, sea turtles are being tracked across the world's oceans using satellite tracking devices <<http://www.kustem.edu.my/seatru/satrack/>>.

Monitoring Adelle Penguins in the Antarctic

The Adelle penguin has been identified as an important krill-dependent indicator species and is being used to monitor changes in critical ecosystem components for use in assessment of the conservation of marine living resources in the Antarctic. One project (Southwell and Meyer 2003) is studying the degree to which the feeding range of the penguins overlaps with

the krill fishery in both time and space; variations in the penguins breeding success and food consumption from year to year and the factors responsible; and how much krill can be fished without affecting the penguins that depend on it.

Examples:

- Tracking Adelie penguins around Casey Station to monitor feeding habits (Kerry *et al.* 1999) <http://aadcmaps.aad.gov.au/aadc/metadata/metadata_redirect.cfm?md=AMD/AU/Tracking_SI>;
- Adelie penguin research and monitoring in support of the CCAMLR Ecosystem Monitoring Project Antarctic Science Project No. 2205 <http://cs-db.aad.gov.au/proms/public/report_project_public.cfm?project_no=2205>.

Wandering albatrosses and petrels

Albatrosses wander for thousands of miles around the southern oceans and generally only ever touch land to breed. Little is known of the movements of the different species and individuals – how far they range, where do they over winter, etc. Primary species-occurrence data are being gathered through the use of satellite tracking and observation (Croxall *et al.* 1993).

Examples:

- Platform Terminal Transmitters have been attached to Tasmanian Shy Albatrosses to track albatrosses over a four month period <http://www.wildlifebiz.org/The_Big_Bird_Race/152.asp>;
- Two species of albatross were tracked around Heard Island in the Antarctic <<http://www.aad.gov.au/default.asp?casid=14718>>;
- Satellite tracking of petrels and albatrosses from the tropics to the Antarctic (Catard and Weimerskich 1998).

Impact of Climate Change

Climate change threatens the survival of ecological communities, individual species, and human health and wellbeing. There have been many studies on the impact of climate change on human populations, on roads and dams, island populations, etc. Fewer studies, have examined the impact of climate change on biodiversity, but the use of species-occurrence data in environmental models to examine impacts is increasing, and studies have shown that impact is likely to be considerable. Howden *et al.* (2003), for example, identified impacts on Australia's coral reefs, on rainforests and rangelands, and on the distribution of birds, plants and reptiles. Recent studies have indicated that as many as 18-35% of species will become extinct before 2050 due to climate change (Thomas *et al.* 2004).

On Native Species

The availability of species-occurrence records through distributed systems such as the GBIF Portal has opened up new areas of research, and allows climate change impacts to be studied across ranges of species, climates and regions.

Examples:

- Studies in Australia on the impact of climate change on threatened species, estimated reductions in core climate habitat of between 82 and 84% with 12% of threatened species predicted to become extinct by 2030 (Dexter *et al.* 1995), and even currently

- non-threatened species with limited distributions, or with specific habitat or soil requirements were likely to be significantly impacted (Chapman and Milne 1998);
- Studies in Mexico looked at the impact of climate change on the fauna (Peterson *et al.* 2002a)
<http://www.specifysoftware.org/Informatics/bios/biostownpeterson/Petal_N_2002.pdf>;
 - A study in Brazil looked at the impact of climate change on cerrado species, and examined implications for conservation assessment and reserve selection (Siqueira and Peterson 2003)
<<http://www.biotaneotropica.org.br/v3n2/en/download?article+BN00803022003+item>>;
 - A study of 35 non-migratory European butterflies showed a major shift north in distribution over the past century of from 35-240 km that the authors contributed to global warming (Parmesan *et al.* 1999)
<http://www.biosci.utexas.edu/IB/faculty/parmesan/pubs/Parm_Ntr_99.pdf>;
 - Studies in birds in America has shown a shift in breeding dates in tree swallows (Dunn and Winkler 1999);
 - A study of the adaptation of migratory birds to global climate change was conducted using the European Pied Flycatcher (*Ficedula hypoleuca*) Coppack and Both 2003)
<<http://www.rug.nl/biologie/onderzoek/onderzoekgroepen/dierOecologie/publications/803Pdf.pdf>>.

On Primary Production

Not all climate change is detrimental, and for agriculture, some species will benefit. Other species will grow in places where they have previously been marginal.

Examples:

- In Australia, it is predicted that wheat yield may increase in some areas (Nicholls 1997);
- Studies in Denmark have shown that global climate change is likely to increase yields at high and mid-latitudes (Olesen 2001)
<<http://glwww.dmi.dk/f+u/publikation/dkc-publ/klimabog/CCR-chap-12.pdf>>;
- Research is predicting that different agricultural and forest species will need to be planted in different areas, some areas will require the planting of new varieties, other species will need to be planted earlier, pesticide controls will need to be altered and water regimes may need to be examined
<<http://www.gcrio.org/gwcc/booklet2.html>>.

Desertification

Climate change and desertification are two big issues facing the world. Primary species data are being used as indicators of diversification under climate change

Examples:

- Grassroots indicators for desertification. Experience and Perspectives from Eastern and (Hambly and Angura 1996);
- Global Biodiversity Forum on Linking biodiversity and desertification: a strategic perspective
<http://www.gbf.ch/desc_workshop_old.asp?no=6&app=&lg=EN&now=2>.

- In Cuba, biodiversity data are being used to develop an index of desertification (Negrin *et al.* 2003) <<http://www.unccd.int/actionprogrammes/lac/national/2003/cuba-spa.pdf>>;
- The trialogue of climate change, biodiversity and desertification <<http://www.gdrc.org/uem/Trialogue/trialogue.html>>.

Ecology, Evolution and Genetics

Primary species-occurrence data provides the raw material for revealing patterns, processes, and causes of evolution and ecological phenomena (Krishtalka and Humphrey 2000). The study of vegetation structure and composition is largely dependant on the availability of species-occurrence data. Much of the world's vegetation has been altered in recent centuries and thus the reconstruction of pre-settlement vegetation cover requires a combination of primary species data and modelling against soils, climate, and topography, etc.

Vegetation Classification

The classification and description of vegetation is a first step in understanding the vegetation, its functions and attributes. Primary species-occurrence data are essential for both the classification and description.

Examples:

- Gillison and Carpenter (1994) used functional attributes for the description and analysis of vegetation <http://www.cifor.cgiar.org/publications/pdf_files/WPapers/WP-03n.pdf>;
- VegClass: Vegetation Classification tool <http://www.cifor.cgiar.org/docs/ref/research_tools/vegclass/>;
- UK Habitat Classifications <<http://www.jncc.gov.uk/habitats/habclass/default.htm>>;
- Vegetation Classification Standards (Federal Geographic Data Committee) <http://www.fgdc.gov/standards/status/sub2_1.html>;
- Vegetation of southern Africa <<http://www.plantzafrica.com/vegetation/vegmain.htm>>.

Mapping Vegetation

Vegetation mapping is a key process in understanding the environment, and in providing a context for studying species and their associations. Vegetation mapping covers both the mapping of current vegetation cover as well as interpretation of past vegetation cover in areas that may now be cleared for urbanization, agriculture, etc.

Examples:

- Checklist of Online Vegetation and Plant Distribution Maps (Englander and Hoehn 2004) <<http://www.lib.berkeley.edu/EART/vegmaps.html>>;
- Australian National Vegetation Information System (NVIS) is using species distribution data from herbaria and on-ground survey to prepare a detailed vegetation map for the continent <http://audit.ea.gov.au/ANRA/vegetation/vegetation_frame.cfm?region_type=AUS®ion_code=AUS&info=NVIS_framework>;

- The Australian Natural Resources Atlas v. 2.0 examines native vegetation types and extent in Australia, and looks at what the vegetation was like prior to European settlement
<http://audit.ea.gov.au/ANRA/vegetation/vegetation_frame.cfm?region_type=AUS®ion_code=AUS&info=veg_type>;
- USGS-NPS Vegetation Mapping Program <<http://biology.usgs.gov/npsveg/>>;
- Florida Coastal Everglades LTER Sited – Vegetation Map <<http://fcelter.fiu.edu/maps/>>.

Habitat loss

Habitat loss (including fragmentation) is considered to be one of the largest threats to biodiversity. The study of habitat loss is again dependant upon the availability of species-occurrence data – including data from museums as well as survey.

Examples:

- The study of woodland birds in Australia has shown a major decline as habitat fragmentation increases
<<http://www.wilderness.org.au/campaigns/landclearing/nsw/birdecline/>>;
- Museum collections were used to show a change in proportions between species of small mammals in the prairies of Illinois coincided with habitat destruction (Pergams and Nyberg 2001) <<http://home.comcast.net/~oliver.pergams/ratio.pdf>>;
- A study of tropical forests in the Mbalmayo Forest Reserve in Cameroon, examined species richness for eight groups of animals and compared them with increased disturbance (Lawton *et al.* 1998) <<http://invertebrates.ifas.ufl.edu/LawtonEtal.pdf>>.

Ecosystem function

Ecosystem function describes the way in which ecosystem processes interact internally between its component organisms and externally with the physical environment, and include such processes as nutrient cycling, decomposition, water and energy balance, and flammability. Ecosystem health (Costanza *et al.* 1992) is very dependant on efficient ecosystem function. Many ecosystems around the world are currently undergoing dramatic changes in species composition due to the influence of human activity. These changes often lead to a reduction in species diversity and species richness and to changes in species composition. How these changes affect overall function of the ecosystem and thus its health is the subject of on-going research. This research is very dependant upon the availability of primary species-occurrence data.

Examples:

- The role of biodiversity in ecosystem function (Gillison 2001)
<<http://www.asb.cgiar.org/docs/SLUM%5C05-Ecological%20functions%20of%20biodiversity%5C05-2%20Does%20biodiversity%20play%20a%20significant.ppt>>;
- Biodiversity and ecosystem function online
<<http://www.abdn.ac.uk/ecosystem/bioecofunc/>>;
- BIODEPTH is a program looking at ecosystem functioning in terrestrial herbaceous ecosystems <<http://www.cpb.bio.ic.ac.uk/biodepth/contents.html>>;
- BIOTREE is a long term project looking at tree diversity and function in temperate forests <<http://www.biotree.bgc-jena.mpg.de/mission/index.html>>;

- Soil microbiology is thought to have a key role in efficient ecosystem functioning (Zak *et al.* 2003) <<http://www.bio.psu.edu/ecology/calendar/Zak.pdf>>.

Survey Design - Finding the Gaps

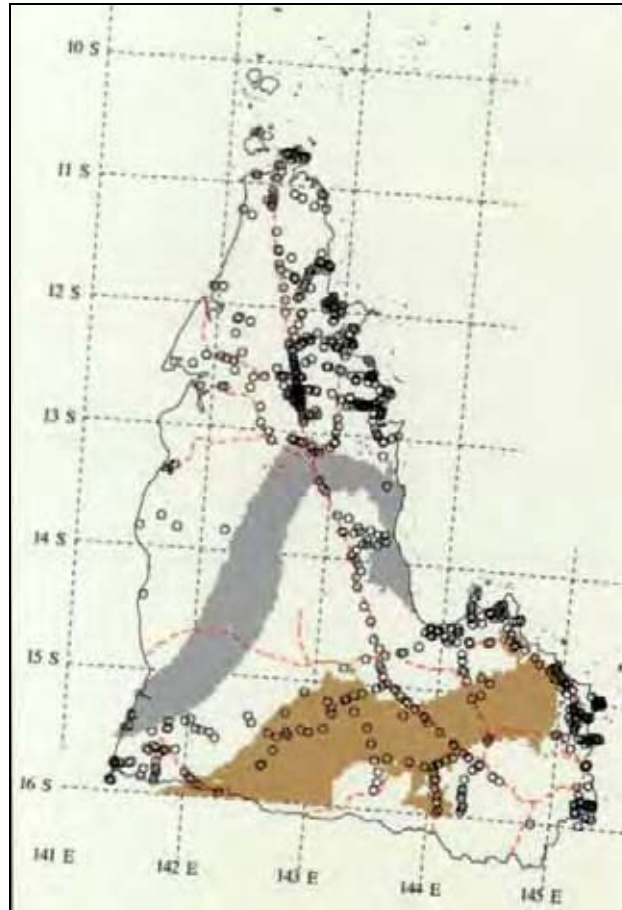


Fig. 7. Environmental regions using climate classes derived from mean annual rainfall and temperature were identified and mapped using GIS. The proportion of biological collections was determined for each class, and surveys planned in areas that were relatively under-surveyed (Neldner *et al.* 1995).

Species occurrence data are a key resource in determining priorities for planning future survey. Although some scientists fear that making their data available electronically will reduce funding support for new biotic surveys and collections (Krishtalka and Humphrey 2000), the opposite is proving to be the case, with increased support for gap filling. By making the data available, geographic, taxonomic and ecological gaps in knowledge are more easily identified, and thus new surveys and survey locations can be planned efficiently and with increased cost-effectiveness (Chapman and Busby 1994).

Examples:

- The U.S. GAP Analysis Program aims at identifying gaps in species conservation <<http://www.gap.uidaho.edu/>>;
- In Australia, environmental and species modelling and biological regionalisation was used to identify key areas of the Cape York Peninsula for further survey (figure 5). A program called VISTR (Visualisation of Taxa, Samples and Regions) was developed (Neldner *et al.* 1995);

- Tour 1 from GBIF Demonstration Project 2003: on the reliability and consistency of Neotropical species distributions can be used to determine appropriate sites for future survey <<http://gbifdemo.utu.fi/>>;
- A BIOCLIM analysis was used in Australia to predict likely habitat for Tarengo Leek Orchid (*Prasophyllum petilum*) based on climatic parameters of the known populations (NSW National Parks and Wildlife Services 2003) <http://www.nationalparks.nsw.gov.au/PDFs/recoveryplan_draft_prasophyllum_petilum.pdf>;
- The South Dakota Gap Analysis program used distributions of native vertebrates to determine survey locations <<http://wfs.sdstate.edu/sdgap/sdgap.htm>>.

Evolution, Extinction and Genetics

Species-occurrence data have been used to study evolution of species, to examine likely species distributions under previous climates, to examine causes of extinctions and to study genetic relationships.

Examples:

- Bioclimatic profiles of a species of *Nothofagus* (*Nothofagus cunninghamii*) were used to estimate Holocene climates in Tasmania (McKenzie and Busby 1992);
- Pollen evidence was used to reconstruct palaeoenvironments in the lower Gordon River valley in Tasmania (Harle *et al.* 1999);
- Species data are used to infer phylogenies <<http://evolution.genetics.washington.edu/book/datasets.html>>;
- The use of Ring species and DNA can infer evolutionary patterns in a range of species <<http://www.origins.tv/darwin/rings.htm>>;
- Studies in Australia are examining the reasons for extinction of the megafauna and the evolution of modern Australian faunal species <<http://science.uniserve.edu.au/school/quests/mgfauna.html>>;
- Evolution and Mass Extinction (Hunt 2001) <<http://cwx.prenhall.com/bookbind/pubbooks/freemanea2/chapter3/custom5/deluxe-content.html>>;
- In Canada, a range of projects is looking at Molecular Systematics and Conservation genetics. Projects include the conservation genetics of endangered species, evolution of unisexuality in reptiles, detection of cryptic species using DNA, etc. <<http://www.rom.on.ca/biodiversity/cbcb/cbmolecu.html>>;
- A study of the evolutionary history of amphibians used molecular data (Feller and Hedges 1998) <<http://evo.bio.psu.edu/hedgeslab/Publications/PDF-files/101.pdf>>;
- The evolution of pattern and mimicry is being studied with butterflies <<http://evo.bio.psu.edu/hedgeslab/Publications/PDF-files/101.pdf>>.

Genomics

Genomics is the study of genes and their functions. Primary species-occurrence data are being used in the study of genomics through frozen tissue collections, such as those at the American Museum of Natural History,

Examples:

- Plant Genome Databases <<http://www.nal.usda.gov/pgdic/>>;
- Institute for Comparative Genomics, American Museum of Natural History <<http://www.amnh.org/science/facilities/hayden.php>>;

- Using genetic data for conservation of the Arabian oryx (Marshall *et al.* 1999) <<http://www.latrobe.edu.au/genetics/staff/sunnucks/homepage/papers/AnimalCons/Marshalletal98.pdf>>;
- Ancient DNA techniques are being used to observe evolutionary processes and to construct phylogenetic trees from fossil bones discovered in the permafrosts of Alaska (Shapiro and Cooper 2003);
- In Finland, adaptive variation is being studied using genomes <<http://cc.oulu.fi/~genetwww/plants/adaptive.html>>;
- DNA bar-coding is being examined for use in biological identifications and conservation (Herbert *et al.* 2003) <<http://barcoding.si.edu>>.

Bioinformatics

In genome terms, bioinformatics includes the development of methods to search databases quickly, to analyse DNA sequence information, and to predict protein sequence and structure.

Examples:

- GenBank Database <<http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html>>;
- EMBL – European Molecular Biology Laboratory <<http://www.embl-heidelberg.de/>>;
- Bioinformatics: Sequence, Structure and Databanks – A Practical Approach (Higgins and Taylor 2000).

Microbial diversity and speciation

James T. Staley²

Since bacteria are the most ancient group of living organisms it is not surprising that the Tree of Life, based on small ribosomal RNA sequence analysis, indicates there are at least 40 kingdoms. Considering this high degree of diversity and the fact that micro-organisms are found in all ecosystems, some of which are extreme environments such as boiling hot springs and acidic habitats at pH 1, it is noteworthy that only about 6,000 species of Bacteria and Archaea have been described and named. One reason for the low number of species is that the species concept used for bacteria is very broad in comparison with that for animals and plants. Scientists are now questioning the microbial species concept not only because of its breadth, but because none of the known bacterial species can be considered endemic to a specific location on Earth. Recently, evidence for endemism has been reported when scientists look at the subspecies level.

Multi-locus sequence analyses of protein genes that are less highly conserved than ribosomal RNA are being used for studies of endemism.

Example:

- One example is that of *Helicobacter pylori* a human pathogen that causes gastric ulcers that may eventually lead to stomach cancer. Using sequence analysis of several protein genes, it has been found that human migration patterns can be discerned by the strains of *H. pylori* that have been harboured in *Homo sapiens* since they dispersed from Africa. Thus, the Maori strains of *H. pylori* contain unique strains that are clearly different from those of European ancestry whose populations migrated to New Zealand more recently. African strains were found in high

² This section was authored by James T. Staley, University of Washington, Seattle, WA, USA.

frequency in West Africa as well as in African Americans. Other patterns have been discerned that can also be explained by human migrations that have occurred in the past 50,000 years (Falush *et al.* 2003);

- Evidence that non-pathogenic bacteria are endemic to hot spring habitats has been recently reported in this newly emerging field. If speciation events are occurring at the subspecies level in micro-organisms, this argues for the need for a redefinition of microbial species. Also, if endemic bacteria exist, this information could be very helpful in forensic studies, because the microbiota on objects removed from an area may contain genetic information about the source of the object;
- The study of speciation is the new revolution in microbiology. Eventually, the numbers of microbial species may exceed many millions.

Archaeological studies

Primary species-occurrence data in the form of fossil collections in museums are used in studying the archaeological history of species.

Examples:

- Researchers at the Illinois State Museum in Springfield are using museum-based fossil data in the scientific literature to plot ranges of North-American mammals over the last 40,000 years on computer-generated maps (Cohn 1995);
- New fossils in Ethiopia open a window on Africa's 'missing years' (Washington University in St. Louis News and Information) <<http://news-info.wustl.edu/news/page/normal/575.html>>;
- African Archaeological Database <http://www.archaeolink.com/african_archaeology.htm>;
- The Age of the Megafauna (Australian Broadcasting Commission) <<http://news-info.wustl.edu/news/page/normal/575.html>>;
- The Zooarchaeology Laboratory Comparative Vertebrate Collection at the Arizona State Museum provides a resource for archaeological studies <http://www.statemuseum.arizona.edu/zooarch/zooarch_browse.asp>.

Environmental Regionalisation

The dividing of an area into regions with similar environmental conditions is possible with the use of species information in conjunction with environmental data and remote-sensing images. Such regionalisations can be used for environmental planning at scales from regional to continental.

National Planning studies

Environmental regionalisations are an extremely valuable tool for planning of conservation and use of natural resources. In Australia, the Interim Biogeographic Regionalisation of Australia (figure 8) is used extensively for conservation planning, sustainable resource management and environmental monitoring.

Examples:

- Interim Biogeographic Regionalisation of Australia (IBRA) was developed using species data, remote sensing data and climate data (Thackway and Cresswell 1995) <<http://www.deh.gov.au/parks/nrs/ibra/version5-1/index.html>>;

- The Australian Government is using bioregions for setting of priority bioregions for developing a national reserve system
<<http://www.deh.gov.au/parks/nrs/ibra/priority.html>>.

Regional Planning Studies

Bioregional planning involves the development of approaches for identifying and characterising regional environmental patterns for use in environmental assessment and planning (Chapman and Busby 1994).

Examples:

- Bioregions are being used in Zimbabwe for conservation planning and for erosion control <<http://www.lancs.ac.uk/fss/politics/people/esrc/pppage2.html>>;
- A New Biogeographic Regionalisation for Tasmania (Peters and Thackway 1998) <<http://www.gisparks.tas.gov.au/dp/newibra/Title&Background.htm>>;
- The Australian Government is using bioregions for integrating conservation and regional planning <<http://www.deh.gov.au/biodiversity/planning/index.html>>. An example is with the Wimmera Catchment Management Authority Pilot Project (Birds Australia 2003) <<http://www.deh.gov.au/biodiversity/publications/wimmera/methods.html>>.

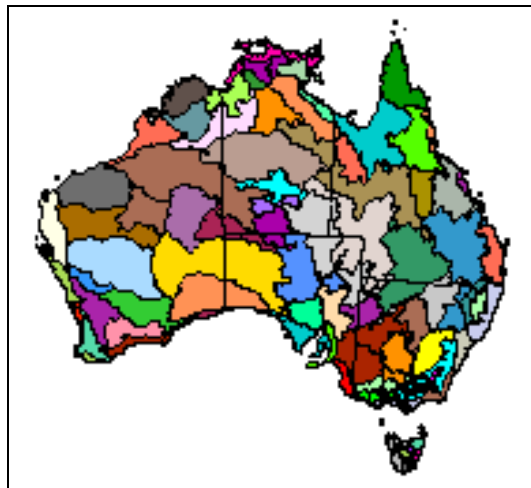


Fig. 8 *The Interim Biogeographic Regionalisation for Australia (IBRA) is a framework for conservation planning and sustainable resource management within a bioregional context. Regions represent a landscape based approach to classifying the land surface using a range of continental data on environmental attributes.*

Marine Regionalisations

Creating meaningful environmental regionalisations of marine areas is not as simple as for terrestrial areas, however they are just as important for conservation planning.

Examples:

- The Interim Marine and Coastal Regionalisation for Australia was created using environmental data such as bathymetry along with species data <<http://www.deh.gov.au/coasts/mpa/nrsmmpa/imcra.html>>;
- Global 200 Ecoregions: Marine <http://www.nationalgeographic.com/wildworld/profiles/g200_marine.html>;

- Canada's National Marine Conservation Areas System Plan
<http://www.pc.gc.ca/progs/amnc-nmca/plan/index_E.asp>.

Aquatic Regionalisations

Aquatic regionalisations are not as common as terrestrial or marine, but are used for managing aquatic ecosystems

Example:

- The management of aquatic ecosystems using macroinvertebrate regionalisations (Wells *et al.* 2002).

Conservation Planning

In order to conserve biodiversity in a long-term sustainable manner, it is important to use species-occurrence data to determine conservation priorities. It is not possible to preserve all populations of all species on earth (Margules *et al.* 2002). It is not even possible to conserve representatives of all species in traditional reserves. Biodiversity has only recently become the most important consideration in reserve selection. Key elements of the priority setting process are complementarity, replication, representativeness and irreplaceability.

Gaston and others (2002) identified six distinct phases in the conservation planning process. The first of these is the compiling of data on biodiversity, reviewing existing data, collecting new data where time and resources permit, and collecting details on locations of threatened and other priority species in the region. The data are an essential first step, and none of the other processes will or can operate without the relevant data.

Rapid Biodiversity Assessment

Most rapid biodiversity assessment projects require extensive amounts of species-occurrence data in order to come up with a meaningful result. Such projects have been very expensive, and the collection of data, and especially species-occurrence data, has been the most time-consuming aspect of these projects (Nix *et al.* 2000)

Examples:

- The BioRap biodiversity assessment and planning study of Papua New Guinea
<<http://www.amonline.net.au/systematics/faith5a.htm>>;
- Papua New Guinea Country Study on Biological Diversity (Sekhran and Miller 1995);
- Amazonian Biodiversity Estimation
<http://www.amazonia.org/SustainableDevelopment/Jauaperi/Biodiversity/ALMA/ABDE/ABDE_3.htm>;
- Rapid biodiversity surveys in Indonesia
<http://www.opwall.com/Indonesia_biodiversity_surveys.htm>;
- Rapid Ecological Assessment in the Spanish Creek Wildlife Sanctuary in Belize
<http://biological-diversity.info/Spanish_Creek.htm>.

Identifying Biodiversity Priority Areas

Biodiversity conservation planning and assessment requires the identification of areas that represent the biological diversity of a region, country or biome (Margules and Redhead

1995). Setting priorities involves deciding what biodiversity to conserve and how much of each species, etc.

Examples:

- Tools for Assessing Biodiversity Priority Areas (Faith and Nicholls 1996);
- Practical application of biodiversity surrogates and percentage targets for conservation in Papua New Guinea (Faith *et al.* 2001);
- Biodiversity World – conservation assessment using biodiversity modelling <<http://www.bdworld.org/index.php?option=content&task=view&id=16&Itemid=25>>;
- The Biodiversity Toolbox for Local Government is designed to provide councils with the tools, resources and contacts to integrate biodiversity conservation <<http://www.deh.gov.au/biodiversity/toolbox/index.html>>;
- National Land and Water Resources Biodiversity Assessment (Identifying Priorities for Conservation) <http://audit.ea.gov.au/ANRA/vegetation/vegetation_frame.cfm?region_type=AUS®ion_code=AUS&info=bio_asses>;
- Establishing marine priority areas <<http://www.mcabi.org/marineprotected/Marine.htm>>;
- Workshop on priority-setting for biodiversity conservation <http://www.earth.nasa.gov/science/biodiversity/section_d2.html>;
- Papua New Guinea Conservation Needs Assessment (Alcorn 1993).

Reserve Selection

Once biodiversity assessment has been carried out, and priority areas for biodiversity identified, the next step is to select appropriate areas for reserves.

Examples:

- Gap Analysis is used in Idaho in the United States for reserve selection <<http://www.gap.uidaho.edu/Bulletins/9/bulletin9/bulletin9html/aaorsafipobanme.html>>;
- The Australian Museum has a program to examine the use of genetic criteria in reserve selection <http://www.amonline.net.au/evolutionary_biology/research/projects/gcrs.htm>;
- Another project at the Australian Museum is looking at using dung beetles as indicator species to measure and compare genetic diversity and evaluating their use in reserve selection <http://www.amonline.net.au/evolutionary_biology/research/projects/ressel.htm>;
- A study in British Columbia examined sensitivities of reserve selection to decisions about scale, biodiversity data and targets (Warman *et al.* 2004);
- Margules and Pressey (2000) stressed the importance of both off-reserve and on-reserve conservation and the need to manage whole landscapes for production and protection;
- Gap Analysis and reserve selection reference list <<http://www.apec.umn.edu/faculty/spolasky/reserve.PDF>>;
- Pattern analysis allows for environmental representativeness in reserve selection (Belbin 1993);
- Designing protected areas and using critical habitat corridors for giant pandas in China (MacKinnon and De Wulf 1994).

Complementarity

The idea of complementarity is to select a set of conservation areas that together contribute a representation of a maximum number of species (Margules *et al.* 1988). Complementarity is an iterative process – for example if you are wanting every species represented, complementarity chooses the first area with the most species, then it looks for the next area that has the most species not already represented and so on. Species-occurrence data are essential in determining areas using these algorithms.

Examples:

- Complementarity, biodiversity viability analysis and policy-based algorithms for conservation (Faith *et al.* 2003);
- Identifying top priority areas as those that make the highest contribution to a representative complementary set (Faith and Walker 1997);
- A new database on the distribution of vertebrate species in a tropical continent allows new insights into priorities for conservation across Africa (Brooks 2001);
- In Oregon, reserve-selection algorithms were compared using terrestrial vertebrate data (Csuti *et al.* 1997);
- A recent whole-country planning study for Papua New Guinea illustrated the importance of complementarity-based trade-offs in determining conservation priorities (Faith and Walker 1996) <<http://www.ias.ac.in/jbiosci/jul2002/393.pdf>>.

Ex-situ Conservation

Not all biodiversity conservation can occur in formally established conservation reserves. Off-reserve or ex-situ conservation is also important and zoological and botanical gardens play an important role in the conservation of rare and threatened species and in captive breeding programs. Species-occurrence data are essential sources of information for institutions and individuals running ex-situ conservation programs.

Zoological Gardens

Zoos now play a major role in conservation of rare species. Many zoos have captive breeding programs, and some are being used to breed up populations of rare species for release back into the wild.

Examples:

- The Przewalski horse is being bred in zoos around the world for release back into the wild <<http://www.imh.org/imh/bw/prz.html>>;
- The IUCN is backing the captive breeding of foxes, wolves, jackals and dogs for reintroduction to the wild <<http://www.canids.org/1990CAP/10captvb.htm>>;
- Reproductive tissue from endangered animals is being preserved in Australia for future breeding programs <http://www.monash.edu.au/pubs/eureka/Eureka_95/freeze.html>;
- In 1995, zoological institutions around the world developed the World Zoo Conservation Strategy <<http://www.zoo.nsw.gov.au/content/view.asp?id=47>>.

Botanical Gardens

Botanical gardens play a similar role to zoos, but with plants. Many rare plants are grown and bred for release to nurseries, thus releasing pressure on wild populations, some species are reintroduced into the wild, and others are conserved in the gardens themselves.

Examples:

- The Green Legacy - botanical gardens and conservation in Canada <http://www.rbg.ca/greenlegacy/pages/botanical_pg2.html>;
- The growing of rare plants in Australian botanical gardens <<http://www.anbg.gov.au/chabg/bg-dir/collections.html>>;
- The Weight of a Petal: The Value of Botanic Gardens (Bruce Rinker) <<http://www.actionbioscience.org/biodiversity/rinker2.html>>;
- A Handbook for Botanic Gardens on Reintroductions of Plants to the Wild (Akeroyd and Wyse-Jackson 1995);
- A Reference List for Plant Re-Introductions, Recovery Plans and Restoration Programmes (Royal Botanic Gardens Kew) <<http://www.rbgkew.org.uk/conservation/reintro.html>>;
- The location of the Wollomi Pine (*Wollemi nobilis*) in Australia was kept secret while botanic gardens grew stock for distribution to nurseries to reduce pressure on wild stocks <<http://home.bluepin.net.au/yallaroo/conservationandcult.htm>>.

Wildlife parks

Wildlife parks – both zoological and botanical – are another place where ex-situ conservation is occurring.

Examples:

- The South Lakes Wild Animal Park in the UK has a large conservation program <<http://www.wildanimalpark.co.uk/>>;
- San Diego Zoo's Wild Animal Park also has some major conservation programs <<http://www.sandiegozoo.org/conservation/zooprojects.html>>;
- Cleland Conservation Park in South Australia aims to conserve both animals and plants <<http://www.environment.sa.gov.au/parks/cleland/>>;
- Many private sanctuaries are being established for the preservation of plants and animals <<http://www.environment.sa.gov.au/biodiversity/sanctuary.html>>.

Sustainable Use

There is an increasing move towards mixing conservation and sustainable use. Not all countries can lock up land in traditional conservation reserves, and are developing sustainable use areas utilising local communities and biodiversity data.

Examples:

- In South Africa, the Ezemvelo Nature Reserve is being proposed as a economically independent conservation-based reserve that utilises its natural resources in a sustainable manner (Sonnekus and Breytenbach 2001);
- In Costa Rica, the Guanacaste Conservation Area has been set up as a sustainable-use reserve with the support of the local community (Janzen 1998, 2000);
- The United Nations Man and the Biosphere Programme aims to reconcile the conservation of biodiversity with its sustainable use <<http://www.unesco.org/mab/>>.

Seed Banks and Germplasm Banks

The conservation of biodiversity through the long-term storage and preservation of seeds and germplasm is another use to which species data is being put.

Examples:

- Millenium Seed Bank Project is a global collaborative project to safeguard plant species from extinction <<http://www.kew.org/msbp/>>;
- The Chinese Academy of Sciences is developing a germplasm bank for wildlife of SW China <<http://english.cas.ac.cn/english/news/detailnewsb.asp?infoNo=24630>>;
- GenBank Database <<http://www.psc.edu/general/software/packages/genbank/genbank.html>>.

Natural Resource Management

Improved information on biodiversity will enhance the ability of resource managers to identify areas of high species diversity, high endemism, and exploitable resources, and improve efforts at protecting and managing natural resources. (Page *et al.* 2004)

Land Resources

The need for management of land resources in a sustainable manner is becoming recognised as an increasingly important issue. The increasing wealth of high-resolution biodiversity data is essential for land use planning and management decisions.

Examples:

- Natural resource management and vegetation – an overview – Australia. <http://audit.ea.gov.au/ANRA/vegetation/vegetation_frame.cfm?region_type=AUS®ion_code=AUS&info=NRMV_overview>;
- Regional Land Use Plans and Land Resource Management Plans (LRMPs) in British Columbia <<http://srmwww.gov.bc.ca/rmd/lrmp/>>;
- In Cuba, biodiversity data are being used to fight against desertification (Negrin *et al.* 2003) <<http://www.unccd.int/actionprogrammes/lac/national/2003/cuba-spa.pdf>>;
- Managing Natural resources in Africa and the Middle East <http://web.idrc.ca/en/ev-3313-201-1-DO_TOPIC.html>;
- The IUCN Sustainable Use site <<http://www.iucn.org/themes/sustainableuse/>>;
- South African Institute for Natural Resources <<http://www.inr.unp.ac.za/>>.

Water Resources

Water resource management, involves sustainable management and use including the development of water quality indicators and the biological control of weeds.

Examples:

- Population growth (with demands for agriculture and hydroelectric power) is combining with climate change to create water stress in Africa (Schultze *et al.* 2001);
- The World Bank – Water Resource Management site <<http://lnweb18.worldbank.org/ESSD/ardext.nsf/18ByDocName/WaterResourcesManagement>>;
- US-China Water Resource Management Program <<http://www.lanl.gov/projects/chinawater/main.html>>;
- Macroinvertebrates are used as indicators of water quality (Maryland Department of Natural Resources)

<http://www.dnr.state.md.us/streams/pubs/freshwater.html#Where%20and%20when%20are%20freshwater%20benthic>>;

- The U.S. EPA water quality and aquatic biology program
<<http://www.epa.qld.gov.au/register/p00736ad.pdf>>.

Environment Protection

Environment protection covers a broad area, and is mostly thought of as protecting the environment from human-induced pollution. But it is much broader than that, and involves protecting the environment from all forms of human-induced impact such as climate change, impacts of the built environment on the natural environment, etc.

Examples:

- Australia's Environment Protection legislation uses an on-line decision support system to monitor impacts of development, agriculture and fishing, etc. on matters of environmental significance such as World Heritage sites, threatened and migratory species, important wetlands. Primary species-occurrence data are a major source of background information for the decision support system (Chapman *et al.* 2001)
<<http://www.deh.gov.au/erin/ert/epbc/index.html>>;
- The US Environment Protection Authority uses species-occurrence data for many aspects of environment protection <<http://www.epa.gov/>>.

Environmental Monitoring

Monitoring of the environment through time is an often-neglected issue, but one that is essential for continual management of environmental resources.

Examples:

- Long-term Monitoring of Australia's Biological Resources (Redhead *et al.* 1994);
- Environmental monitoring in Sweden
<<http://www.svenskamiljonatet.se/cbd/eng/hav/miljoovervakning.htm>>;
- University of Waterloo students collect data in forest biodiversity plots every summer as part of a third year course in Environmental Monitoring
<http://www.escarpment.org/Monitoring/mon_forestbio.htm>;
- The Albufera International Biodiversity Group (TAIB) uses volunteers to collect data for monitoring environmental change
<http://www.medwetcoast.com/article.php3?id_article=200>;
- Birdlife International uses biodiversity indicators for environmental monitoring
<<http://www.birdlife.net/action/science/indicators/>>.

Agriculture, Forestry, Fisheries and Mining

The fields of agriculture, forestry, fisheries and mining have been among the greatest users of primary species-occurrence data. The identification of appropriate areas for growing crops, the identification of wild relatives of key crop species for genetic breeding, the identification of new species for food, forestry, shelter, fibre and industrial uses, the identification of provenances for use in planting in different areas, the identification of biological control agents for weeds and diseases, the identification of key areas for forestry production and protection, both for plantation and native harvesting, identification and management of

fisheries production, the identification of by-catch, the study of feeding habits, pesticides, contaminants, and the identification of possible mine sites; etc.

Agriculture

A new term, 'agrobiodiversity' or 'agricultural biodiversity', has recently been defined by Decision V/5 of the Fifth Conference of the Parties to the Convention on Biological Diversity, as including "... all components of biological diversity of relevance to food and agriculture, and all components of biological diversity that constitute the agro-ecosystem" (<http://www.biodiv.org>). This includes ecological services such as nutrient cycling, pest and disease regulation (natural biological control), pollination, wildlife habitats, hydrological cycles, carbon sequestration, and climate regulation as well as cultural aspects, including tourism (Miller and Rogo 2001).

The food industry in the United States alone is estimated to be worth \$800 billion per year (Pimental *et al.* 1999). All of this is based on biological species whether they are plants such as corn, wheat, rice, soybeans, or other food crops, animals like cows, pigs and poultry, or fungi such as mushrooms. Biological species are also used in the agricultural industry for landscape restoration, biological pest control, sport, pets and food processing. Primary species-occurrence databases are a key source of information for use by agriculturalists.

New crops and wild relatives

The world is always looking out for new species for use in agriculture. Primary species databases are being used to identify wild relatives of species currently being used for agriculture, or new species that may have been used by indigenous peoples. In addition, wild relatives of cultivated crops are being examined for genetic transfer to control weeds, improve growth rates, reduce water use, etc.

Examples:

- Close relatives of cultivated rice, including *Oryza rufipogon*, *O. nivara*, *O. longistaminata*, and *O. glumaepatula* are commonly found or coexist in rice farming systems of many Asian, African, and American countries. The use of these species in cross breeding has been in practice for hundreds of years, and more recently biotechnology has been used to transfer specific genes to increase levels of beta-carotene, protein content, disease and insect resistance, herbicide resistance, and salt tolerance (Lu 2004);
- In Brazil, controlled and natural hybridisation is occurring between cassava (*Manihot esculenta*) and its wild relatives. Studies are being carried out to look for, or breed, new hybrids for improved production and fertility (Nassar 2003) <http://www.funpecrp.com.br/gmr/year2003/vol4-2/gmr0047_full_text.htm>;
- The Desert Quandong (*Santalalum acuminatum*) is a plant traditionally used by Australian aborigines. It is now being developed as a commercial food source <<http://sres.anu.edu.au/associated/fpt/nwfp/quandong/Quandong.html>>.

Provenances and wild relatives

The identification of new provenances of cultivated species has been a tradition going back many hundreds of years. Primary species-occurrence databases can now help in that search as point records are extracted over the Internet, allowing the identification of new populations and areas for study.

Examples:

- In New Zealand, four new provenances of the plant *Cordyline australis* were selected to examine their suitability for fructose production (Harris 1994);
- In Australia, suitable provenances of species of *Acacia* are being sought for use for grazing livestock (Dynes and Schlink 2002);
- In Central Africa, germplasm of eighty-five provenances of Eru (*Gnetum africanum* and *Gnetum buchholzianum*), species that are valued as highly nutritious green vegetables, have been selected for genetic improvement and ex-situ cultivation and management (Shiembo 2002) <<http://www.fao.org/docrep/X2161E/x2161e06.htm>>;
- Potential for Seed Gum production in *Cassia brewsteri* (Cunningham *et al.* 2001) <<http://www.rirdc.gov.au/reports/NPP/UCQ-12A.pdf>>;
- Seeds for Success program in the United States is collecting seeds of species for use in stabilisation, rehabilitation and restoration of degraded land <<http://www.nps.gov/plants/sos/>>.

Food processing

The use of species in food production goes back many thousands of years with the use of yeasts in the production of alcohols and bread, and bacteria for cheese production. For example, the many varieties and flavours of wine come from the extensive selection of possible grapes to be fermented and from the vast array of yeasts and bacteria that may be used. Winemakers and beer brewers are always on the lookout for new and improved yeast varieties.

Examples:

- The Role of Yeast in Production of Alcoholic Beverages. <<http://www.botany.hawaii.edu/faculty/wong/BOT135/Lect14.htm>>;
- Bacteria are used in the production and processing of sour creams, buttermilk, yoghurt, cheese, sauerkraut, pickled vegetables, chocolate, coffee, vinegar, etc. and manufacturers are always on the lookout for new and improved species to use to introduce new flavours and products <<http://www.bacteriamuseum.org/niches/foodsafety/goodfood.shtml>>.

Harvesting of wild populations

The harvesting of wild populations of plants and animals for food and ornament is another major industry that benefits from the availability of species-occurrence data and databases. The harvesting of native animals is a controversial subject, but it is an industry that is important in many developing countries. The use for forestry is considered elsewhere, but the harvesting of flowers from wild areas is a large industry in countries like South Africa and Australia. Species-occurrence data are used in the identification of species suitable for harvesting, and for determination of areas with sustainable populations.

Examples:

- Wildlife harvesting in the Fynbos area of South Africa provides income for 20,000 people (Lee 1997) <<http://www.ars.usda.gov/is/pr/1997/971010.2.htm>>;
- Some South American fruits and yams are grown under “semi-wild” cultivation, for example *Spondias mombin* (Campbell 1996) <<http://www.hort.purdue.edu/newcrop/proceedings1996/V3-431.html>>;
- Twenty two species of animal are harvested from the wild in Africa (Ntiamao-Baidu 1997) <<http://www.fao.org/docrep/W7540E/w7540e00.htm>>;
- In Brazil, many native fruits are used for flavouring ice-creams and as fruit juices <http://www.maria-brazil.org/brazilian_sherbets.htm>.

Beneficial Insects in agriculture

As well as their huge detrimental impact on agriculture, insects are also an important positive contributor.

Examples:

- Honey industry profile (Saskatchewan Agriculture, Food and Rural Revitalization) <<http://www.agr.gov.sk.ca/docs/crops/apiculture/HoneyIndustry.pdf>>;
- Silk Business in Iran <http://www.iccim.org/English/Magazine/iran_commerce/no1_1999/17.htm>;
- The economics of apiculture and sericulture modules for income generation in Africa (Raina 2000);
- Cash crops (e.g. butterflies or chemical extraction), mini-livestock (Odhiambo 1977);
- Termite nests are also used for building materials (Swaney 1999: 435);
- Species-occurrence data was used to improve pollination in Oil Palms in Malaysia <http://www.bionet-intl.org/case_studies/case14.htm>.

Weeds and Pests

The financial impact of weeds, pests and diseases on agricultural production is enormous (Suarez and Tsutsui 2004). Species that generally cause the greatest impact are introduced from other areas (Pimental *et al.* 1999), and these are covered separately under the section on invasive species, above. Not all pests and diseases are introduced, however, and their identification, control and management can be an important issue for farmers. Weeds for example can provide valuable food resources for pollinating insects. Often, past clearing of land for agriculture has meant increased grassland for grazing animals and seed-eating birds such as Kangaroos and Corellas in Australia. Primary species-occurrence databases can be important in the identification of weeds and pests of agriculture and for studying their distributions.

Examples:

- Some animals have adapted well to the changed landscape of Australia and their numbers continue to increase. These include western grey kangaroos (*Macropus fuliginosus*), galahs (*Cacatua roseicapilla*), ravens (*Corvus coronoides*), Australian magpies (*Gymnorhina dorsalis*), corellas (*Cacatua tenuirostris*) and the Port Lincoln Parrot (*Barnardius zonarius*). Other, related species can be very rare, so identification is important for their management (Hindmarsh 2003). <http://portal.environment.wa.gov.au/pls/portal/docs/PAGE/DOE_ADMIN/TECH_REPORTS_REPOSITORY/TAB1019581/WRM33.PDF>;
- Only 5 of 43 species of macropods (kangaroos) in Australia can be harvested, and counts are made every year to determine the numbers allowed for harvesting. Identification is important so as to get accurate numbers and so that more threatened species aren't killed by mistake <<http://www.dfat.gov.au/facts/kangaroos.html>>;
- Correct identification of a fungus of wheat in the USA saved \$5 billion/year in wheat exports <http://www.bionet-intl.org/case_studies/case8.htm>.

Invertebrate pests

Invertebrate pests, especially insects, cause massive losses to production every year, and are a major cause of famine (plague locusts) in many parts of Africa and elsewhere. The identification of pests is another role where species data plays a crucial role.

Example:

- The National Centre for Integrated Pest Management in India, has developed a program to map the geographical distribution of all pests of major crops in the country <<http://www.ncipm.org.in/Maps.htm>>;
- The International Development Research Centre is setting up an insect identification and biosystematic service for agriculture Africa south of the Sahara <http://web.idrc.ca/en/ev-26155-201_870175-1-IDRC_ADM_INFO.html> ;
- Intercropping increases parasitism of pests (Khan *et al.* 1997).

Plant and animal pathogens

There are an estimated 50,000 parasitic and non-parasite diseases of plants in the United States alone, most of which are caused by fungus species. Mycological species databases, including living collections, can be important for the identification and control of many of these species.

Examples:

- The Ecological Database of the World's Insect Pathogens offers information on fungi, viruses, protozoa, mollicutes, nematodes, and bacteria that are infectious in insects, mites, and related arthropods <<http://cricket.inhs.uiuc.edu/edwipweb/edwipabout.htm>>;
- With links to primary species-occurrence databases, Kansas State University uses geographic tools to track plant pathogens <http://www.innovations-report.de/html/berichte/agrar_forstwissenschaften/bericht-27646.html>;
- Modelling the spatial distribution of important South African plantation forestry pathogens (van Staden *et al.* 2004).

Forestry

The forestry industry is an enormous industry around the world. It is an industry that has traditionally utilised native and wild populations, but one that is gradually moving toward plantation forestry. Primary species-occurrence data play a role in both areas, firstly through identifying species and areas for forest production, and attempting to balance that with conservation, and secondly in determining what species and provenances will most suitably grow where.

Balancing forestry and conservation

Native forest industries rely on species distribution data to find locations of new species and areas for forest production. Species-occurrence data are also used to develop sustainable forestry management processes through setting aside restricted areas for native harvest, and using methods described elsewhere in this paper (see Conservation Assessment) for determining those areas to be set aside for conservation.

Examples:

- National Indigenous Forest Inventory for South Africa (Wannenburgh and Mabena 2002) <<http://www.dwaf.gov.za/Forestry/FTIS/symp2002/inventory.doc>>;
- The National Forestry Programme for Swaziland examines biodiversity values and multi-uses of forestry land <http://www.ecs.co.sz/forest_policy/fap/index.htm>;
- Regional Forest Agreements in Australia. Government of Tasmania & the Commonwealth of Australia

<http://www.affa.gov.au/content/output.cfm?ObjectID=89389274-95D8-4380-BD9BB177D644820A&contType=outputs>>;

- Using process-based and empirical forest models in eucalypt plantations in Brazil (Almeida *et al.* 2003);
- Studies in Australia use species-occurrence data in modelling and conservation assessment to balance forestry and biodiversity (Faith *et al.* 1996).

Plantation forestry

The use of plantation forestry is increasing throughout the world, and techniques are being used to determine the most suitable location for species to be grown. Species-occurrence data are linked with environmental modelling to determine climate profiles from native areas and then applying those profiles to areas and countries where the plantation is to be grown.

Examples:

- Matching Trees and Sites using environmental modelling (Booth 1996)
- Modelling Forest Systems. This book looks at forest models, tools and approaches to forest modelling, including distribution modelling – some of it using species-occurrence data. (Amaro and Soares 2003).

Provenance identification

The selection of the most appropriate provenance of a species to grow in a new plantation area is extremely important. The selection can not only use present-day conditions, but can model and select for future climate conditions, etc.

Examples:

- Selecting species and provenances of Australian trees for growing in Australia, China, Thailand, Laos, Cambodia, Vietnam, Indonesia, Philippines and Zimbabwe, as well as regions such as Southeast Asia, Africa, and Latin America (CSIRO Australia) <<http://www.ffp.csiro.au/pff/species/>>;
- Matching Trees and Sites using environmental modelling examines the provenances of tree species from Australia for planting in China and South-east Asia (Booth 1996);
- In India, new provenances of the genus *Leucaena* are being sought with the aim of finding provenances that can introduce straighter stems, later flowering and lower seed set. <<http://www.forests.qld.gov.au/resadv/research/qficonf/qfri6.htm>>;
- In Vietnam, *Acacia* species and provenances are being selected for large-scale plantings (Ngia and Kha 1996). Between 1982 and 1995, 18 species and 73 provenances from 5 species of *Acacia* were trialled at 8 localities across Vietnam <<http://www.forests.qld.gov.au/resadv/research/qficonf/qfri6.htm>>;
- Climate change studies in the UK, have concluded that new provenances of existing species will need to be found in order that new plantations may be adapted to the warmer, and possibly drier conditions expected in the future (Cannell *et al.* 1989).

Fishing

Fishing and fisheries are an important industry and user of species distribution data. With ever increasing pressure on fishing stocks as evidenced by the decline in stocks of Cod fish in the Northern Atlantic (Crosbie 1992, Meisenheimer 1998). Being able to track stocks and movement of fish throughout marine and fresh waters is essential to the long-term sustainable management of commercial fish stocks. The identification of species caught in by-catch is also important for conservation and resource management.

Resource management

Resource management of both marine and fresh-water fisheries is become a critical issue around the world. A large proportion of the world's coastal population is almost entirely dependant on the fishing industry for their livelihood. The use of distributed data and information to make important resource decisions is becoming increasingly important.

Examples:

- The Gulf of Maine Biogeographic Information System is developing a methodological framework for accessing and distributing marine biogeographic data. The system will provide information and tools for better understanding and regulating fish populations (Tsontos and Kiefer 2000)
<<http://gmbis.marinebiodiversity.ca/aconw95/aconscripts/gmbis.html>>;
- The US National Marine Fisheries Service provides automated data summaries of US commercial fisheries landings for fish and shellfish. The volume and value of landings from 1950-2002 landings can be summarized by: years, states and species
<<http://www.st.nmfs.gov/st1/commercial/>>;
- FAO Species Identification and Data Programme (SIDP)
<www.fao.org/fi/sidp/products.htm>;
- Studies in the Bering Sea have examined long-term oceanic primary production and ecosystem changes and shown significant declines in productivity by as much as 25-45% between 1947 and 1997 (Schell 2000)
<http://www.alaskasealife.org/documents/Education/Teacher_guide.pdf>
- The identification of marine hotspots in New Zealand, allows for protection of valuable spawning grounds <http://www.bionet-intl.org/case_studies/case25.htm>.

Overfishing

Overfishing of native stocks is becoming an ever-increasing issue. Overfishing of cod in the northern Atlantic has caused major disruption of whole populations of people, for example on Newfoundland where people have had to find new occupations. Species-occurrence data are used for monitoring stocks.

Examples:

- Closure of cod fishing Action Plan (CNLBSC 2003)
<http://www.cbsc.org/nf/search/display.cfm?Code=6145&coll=NF_PROVBIS_E>;
- What is the problem with cod? (Meisenheimer 1998)
<<http://www.imma.org/codvideo/whatproblemcod.html>>;
- A study of the effects of fishing on deep-water fish species to the west of Britain was carried out in the 1970s and 1980s (Basson *et al.* 2002).

Freshwater

Commercial freshwater fisheries are also important in many parts of the world. In many countries, freshwater fishing is largely recreational, but commercial fishing is still an issue in those countries, as well as in countries with large inland lakes, and large inland fishing industries.

Examples:

- Freshwater Fisheries Management Policy of the Victorian Government in Australia
<<http://www.nativefish.asn.au/fwpolicy.html>>;
- Fish and Fisheries of the Great Lakes Region, Canada with information on species, ecology, etc.<<http://www.great-lakes.net/envt/flora-fauna/wildlife/fish.html>>;
- "Farming Freshwater Prawns" is an FAO technical paper that examines nomenclature and distribution as well as providing a manual for culture of the Giant

River Prawn (*Macrobrachium rosenbergii*).

<<http://www.fao.org/DOCREP/005/Y4100E/y4100e00.htm#TOC>>;

- Inland capture fisheries and enhancement: status, constraints and prospects for food security (Coates 1995).

Bycatch

The identification and reduction of bycatch from commercial fishing is becoming an international issue as more and more marine species are becoming endangered. The monitoring of bycatch has become a requirement of some governments, and methods to reduce the numbers of species and amount of bycatch have been put in place.

Examples:

- A program in the Gulf of Mexico looks at the effects of bycatch on the conservation of fisheries resources in the Gulf (Burrage *et al.* 1997).
<<http://www.rsca.org/docs/ib324.htm>>;
- Tuna Bycatch Action Plan stresses the need for correct identification of turtles, and the need for species identification posters and booklets.
<http://www.afma.gov.au/fisheries/etbf/mac/mac54/item3_2.pdf>;
- The CSIRO Fact Sheet on Conserving Australian Sharks and Rays also stresses the need for “identification guides to assist in the collection of comprehensive information on bycatch species to underpin sustainable management”
<<http://www.marine.csiro.au/LeafletsFolder/53guide/53.html>>.

Contaminants

The identification of contaminants in fish and their monitoring through time to determine suitability for human consumption is another use for species-occurrence data. Fish are also good organisms for testing of water quality through the accumulation of toxins.

Examples:

- Testing for Persistent Environmental Contaminants in Fish and Wildlife (Schmitt and Bunck 1995);
- Integrated Fish Monitoring in Sweden (Sandström *et al.* 2004);
- Use of fish specimens from Richter Museum to analyse historic DDT levels in avian food webs <<http://www.uwgb.edu/davisj/biodiv/richter/resources.htm>>;
- National Contaminant Biomonitoring Program of the USGS studies concentrations of arsenic, cadmium, copper, lead, mercury, selenium, and zinc in freshwater fishes of the United States <<http://www.cerc.cr.usgs.gov/data/ncbp/ncbp.html>>.

Nursery and Pet Industry

Plant nurseries

The Nursery industry is a large user of species names and thus benefits greatly from the use of species-occurrence data. Nurseries are always looking for the names of the plants they sell, and information on their distributions for adding to the labels.

Examples:

- The Society for Growing Australian Plants tracks name changes for informing growers and nurseries that sell Australian plant species
<<http://farrer.csu.edu.au/ASGAP/changes.html>>;
- The Ornamental Plants database provides details on hundreds of cultivated plants with names and information
<<http://www.msue.msu.edu/msue/imp/modzz/masterzz.html>>.

Orchids and mycorrhiza

The cultivation of many terrestrial orchids requires an association with specific mycorrhiza and species databases can assist with the identification of these associations.

Examples:

- Many studies have been carried out at the Australian National Botanic Gardens on the symbiotic germination of terrestrial orchid species (Clements and Ellyard 1979) <<http://www.anbg.gov.au/cpbr/summer-scholarship/2003-4-offer-clements.html>>;
- In Costa Rica, studies on the relationship of mycorrhiza and orchid cultivation are being carried out at the Lankester Botanical Gardens (Rivas *et al.* 1998).

Pets

The pet industry is a huge industry world wide. Pet shops, etc. require information on the names and original localities of many of the animals they sell.

Examples:

- In the US alone, 19 million birds live as household pets <<http://www.birdsnways.com/>>;
- Index of exotic pets <<http://exoticpets.about.com/cs/resourcesgeneral/a/exoticpetsatoz.htm>>.

Mining

The mining industry would seem to be an unlikely user of species-occurrence data, but there are two major areas of biodiversity use in the mining industry. Some species are indicators of high mineral concentrations and are even used in mining in some rare cases; others are used in mine site regeneration.

Examples:

- *Terminalia alata* is used in India to indicate Copper mineralisation (Pujari and Shrivastava 2001);
- Phyto-mining is the use of plants to extract valuable heavy-metal minerals from soils <<http://www.ars.usda.gov/is/pr/2000/000622.htm>>;
- Phyto-mining of gold in New Zealand and Brazil <http://www.gold.org/discover/sci_indu/gold2003/pdf/s36a1355p976.pdf>;
- Phytoremediation uses plants to clean up soils <<http://www.ars.usda.gov/is/AR/archive/jun00/soil0600.htm>>;
- The influence of insects on soil chemistry may even be utilised in prospecting for minerals (e.g. Watson 1974);
- Rehabilitation of mines and other disturbed sites <<http://www.otago.ac.nz/geology/features/restoration/wangaloe/wangaloe.html>>;
- Species of the genus *Polycarpea* have been used to indicate copper as they generally only grow on copper rich soils (Nicholls *et al.* 1965).

Mining and waste

Species-occurrence data are being used for biotechnology uses such as mining and pollution monitoring and control.

Examples:

- Bacteria are being used to clean up toxic waste sites including nuclear sites <<http://sfgate.com/cgi-bin/article.cgi?f=/c/a/2003/07/14/MN103893.DTL>>;

- Bacteria are being used to extract ore including copper, gold and iron and in waste management leading to cleaner mining technologies
<<http://www.bioteach.ubc.ca/Bioengineering/microbialmining/>>;
- Plants are used as detectors for air pollution and as scavengers of air pollutants (Omasa *et al.* 2002) <<http://www.cplpress.com/contents/C808.htm>>;
- Lichens are used as pollution indicators
<<http://www.earthlife.net/lichens/pollution.html>>.

Health and Public Safety

The importance of species data and their contribution to public health and safety, although increasing in importance, is still largely unknown by the general populace. As mentioned by Suarez and Tsutsui (2004) species-occurrence data “play a critical role in public health and safety as cornerstones in studies of environmental health and epidemiology”. They also play a key role in security through their importance in the prevention, detection, and investigation of various types of bioterrorism (NRC 2003).

Health, both human and environmental, is being impacted upon by climate change along with the recent increase in terrorism and human, animal and plant migrations. Species-occurrence data can contribute valuable insights into the study of pathogens, vectors of diseases, and environmental contaminants (Suarez and Tsutsui 2004). Many diseases (human, animal and plant) are biodiversity-related, and the distribution of both the vectors and the disease agents themselves can be studied using species-occurrence data. When linked to biodiversity modelling programs, the potential spread and rate of spread of some of these species can be predicted, both under present day conditions and under altered climate change regimes, etc.

Diseases and disease vectors

Studies of the West Nile Virus in the Dominican Republic (Komar *et al.* 2003) examined the presence of West Nile virus in bird species and hypothesised possible linkages to migration routes of migratory bird species. The use of distribution modelling (Peterson *et al.* 2003b) successfully tested hypotheses that West Nile virus transmission on large geographic scales was by migratory birds, and using this information in conjunction with a simulation model allowed for new outbreaks and spread to be predicted (Peterson *et al.* 2003b).

Many other viruses are also transmitted by vectors, and entomological collections around the world include many records of mosquitoes which are responsible for the transmission of diseases, including malaria, avian malaria, dengue fever, equine encephalitis, and the already mentioned West Nile virus.

Species-occurrence data have also been used to construct evolutionary histories of viruses in order to develop more robust vaccines (Ferguson and Anderson 2002), to study the origins of HIV (Siddall 1997), to study the origins and track the movement of Avian Influenza (Bird Flu) in native and domestic bird populations (Perkins and Swayne 2002) and studying possible cross-susceptibility of the Rabbit CaliciVirus (RHD) in other animals (Munro and Williams 1994).

In addition, we now have the issue of emerging infectious and parasitic diseases, and the need to document transmission patterns. This cannot be done without species-level identifications of both adults and infective stages (larvae/juveniles) (Brooks and Hoberg 2000).

Examples

- West Nile Virus (Komar *et al.* 2003).
<http://www.specifysoftware.org/Informatics/bios/biostownpeterson/Ketal_EID_2003.pdf>;
- Mosquito-borne diseases (Rutgers University and CDC)
<<http://www.rci.rutgers.edu/~insects/disease.htm>>;
- Rabbit Haemorrhagic Disease (Munro and Williams 1994);
- Origins of HIV (Siddall 1997).

Bioterrorism

A key role in the use of species-occurrence data in controlling terrorism is in tracking the history of infectious diseases and in identifying their sources. Among some of the most important public health-related collections of species-occurrence data are the examples of known viruses and bacteria that are retained and used for comparison with outbreaks of new infections. A recent example of their use was with the anthrax attack on the United States in 2001 where researchers at various Centers for Disease Control and Prevention used specimen collections from the 1960s and 1970s to attempt to identify the anthrax strains used (Hoffmaster *et al.* 2002).

One of the identified challengers to the museum community in the face of national threats of this nature is to be able to provide swift and accurate identifications of possible bioterrorism agents (Page *et al.* 2004).

Examples:

- Anthrax attack on the United States 2001 (Hoffmaster *et al.* 2002);
- Biological terrorism Risk Assessment (University of Kansas, Biodiversity Research Center) <<http://www.specifysoftware.org/Informatics/informaticsbtra/>>.

Biosafety

The flow of genes from modified organisms to their wild relatives is a recognised risk associated with genetically modified crops (Soberón *et al.* 2002). As noted by Soberón *et al.*, the risk is greatest when a crop “spontaneously hybridises with its taxonomically related species”. Species-occurrence data are a necessity if scientists are going to be able to assess these risks by tracking spatial relationships between GMO crops and wild relatives, their potential distributions under various climatic conditions, and the reproductive biology of both groups of plants (Soberón *et al.* 2002).

Examples:

- The Mexican Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (Conabio) (<http://www.conabio.gob.mx/>), is using species-occurrence data obtained from herbaria around the world to study and model potential distributions, and to study the likelihood of genetic transfer (Soberón *et al.* 2002). This information is used several times a week to inform the Mexican Ministry of Agriculture (Soberón *pers. com.* Aug. 2004).

Environmental Contaminants

The monitoring of environmental contaminants in natural populations is another important health-related use for primary species-occurrence data. An example is the use of the Swedish Museum of Natural History’s Environmental Specimen Bank to monitor contaminants in

faunal species and to study the effects of noxious substances on endangered and threatened species. Another example is the tracking of pesticides, fungicides, etc. in streams through studying contaminants in populations of native amphibia. In conservation studies on the Californian Condor (*Gymnogyps californianus*), it was found that contamination with lead (and possibly DDT) were major causes of its decline towards extinction through increased mortality (Janssen *et al.* 1986). Museum collections have been used to examine lead and DDT levels through both time and space (Ratcliff 1967). Other studies have looked at increasing mercury levels in marine ecosystems by examining mercury levels in the feathers of seabirds breeding in various areas of the world and compare the levels achieved with historic specimens from the same localities held in natural history museums (Monteiro and Furness 1998, Thompson *et al.* 1998). Birds accumulate heavy metals from their food and secrete them into their growing feathers during molt (Green and Scharlemann 2003). Long-term changes and spatial variation in heavy metal concentrations can easily be studied using such collections.

Examples:

- Environmental Specimen Bank (Swedish Museum of Natural History) <<http://www.nrm.se/mg/mpb.html.en>>;
- Environmental Contaminants of Amphibians in Canada (*Froglog* 16: 1996) <<http://www.open.ac.uk/daptf/froglog/FROGLOG-16-5.html>>;
- Mercury in Feathers from Birds of the Southeastern Pacific: Influence of Location and Taxonomic Affiliation <http://cars.er.usgs.gov/posters/Ecotoxicology/Mercury_in_Bird_Feathers/mercury_in_bird_feathers.html>.

Antivenoms

Snakebite and spider bite are common in many parts of the world, and nowhere more so than Australia where more than 3000 cases are reported annually (Queensland Museum 2004). Many of the world's most venomous snakes are found in Australia. The accurate identification of a snake responsible for a bite allows for the correct antivenom to be administered. Species-occurrence data can limit the areas for which specific antivenoms may need to be stored, and assist in quicker identification of the snake through geographic sifting. This can be important from both from a health point of view, and because of cost. An ampoule of polyvalent antivenom (a cocktail of separate antivenoms) costs \$1600 in Australia, compared to \$300 to \$800 (depending on the species of snake) for an ampoule of specific antivenom (Queensland Museum 2004). Victims of snakebite may require up to eight ampoules of antivenom, so the cost saving of an accurate identification can be significant, and in addition there are significant health benefits.

Examples:

- Queensland Museum antivenom project <<http://www.qmuseum.qld.gov.au/features/snakes/saving.asp>>.

Parasitology

Parasites are becoming recognised as significant components of the environment and are good models for evolutionary studies (Brooks and Hoberg 2001). Parasites are agents of disease in humans, domestic livestock and native wildlife, and maintain a significant role in ecosystem integrity and stability (Brooks and Hoberg 2000). Parasite collections have traditionally been held in large personal collections and have thus been less available to

researchers than they may otherwise have been (Hoberg 2002). This is now being rectified with new distributed systems like the GBIF Portal. Specimen-based data can serve as historical and temporal baselines for understanding environmental change and human intervention on the distribution of parasites and pathogens (Hoberg 2002).

Examples:

- The United States National Parasite Collection (USNPC) is providing a major resource for systematic, taxonomic, diagnostic ecological and epidemiological research <<http://www.anri.barc.usda.gov/bnpcu/>>;
- The distributions of rodents have been used to study reservoirs and vector sites for a range of parasitic diseases, including Lyme disease – a parasitic disease transmitted to humans via tick bite, Lassa fever in Africa associated with multimammate rats, various hanta viruses in Argentina and Chile (Mills and Childs 1998) <<http://www.cdc.gov/ncidod/eid/vol4no4/mills.htm>>;
- Parasites are being used in studies of evolutionary biology (Dimigian 1999) <http://www.baylorhealth.edu/proceedings/12_3/12_3_dimijian.html>;
- Epidemiology of amoebiasis: an age-old problem solved by taxonomy <http://www.bionet-intl.org/case_studies/case1.htm> .

Safer Herbal Products

Many new herbal medicines are becoming available and being sold through pharmacists and health stores. The safety and purity of these medicines needs to be monitored and tested. To this information on their geographic distribution can be important.

Examples:

- Authentication of Chinese herbal medicines helps to deliver safer medicines <http://www.bionet-intl.org/case_studies/case3.htm> ;
- Testing and standardisation of herbal medicines <<http://www.frhlt-india.org/html/lab.htm#testingmedicines>> .

Bioprospecting

Bioprospecting is the search for, and identification of, plants and animals that may provide products with potential economic value, such as new pharmaceuticals, foods, and other as yet-undiscovered uses. Species-distribution data are needed to assist in determining sites and likely species and requires taxonomic and phylogenetic research and distributional information from natural history collections (Page *et al.* 2004).

Pharmaceuticals

For centuries, plants and animals have been the source of healing products. Today, they are the basis of many of the world's pharmaceutical drug products. Primary species data are used to identify relatives of species that are already sources of active products and to find locations of those and other species for assay.

Examples:

- In Costa Rica, the National Biodiversity Institute (Inbio) is a major player in bioprospecting for pharmaceutical products in the forests of Costa Rica (Janzen *et al.* 1993) <<http://www.inbio.ac.cr/en/>>;

- Natural products research, especially novel chemical aspects of insect-plant interactions and arthropod venoms in Africa (Iwu 1996; Torto & Hassanali 1997; Weiss and Eisner 1998);
- Using plants to produce pharmaceuticals (Council for Biotechnology Information) <<http://whybiotech.ca/canada-english.asp?id=3352>>;
- In Brazil, the FAPESP-Biota program is funding a project to examine plants of the Mata Atlantica (Coastal rainforest) and Cerrado (savannah) for chemical and pharmacological products <<http://www.biota.org.br/projeto/index?show+229>>;
- The Amazon Rainforest is a source for present and future drugs <<http://www.rain-tree.com/>>;
- Ants as a source of pharmaceuticals (Majer *et al.* 2004)
- Plant-derived Drugs: Products, Technology, Applications <http://bcc.ecnext.com/coms2/summary_0002_001960_000000_000000_0002_1>;
- Chemotaxonomy of Xylariaceae uses bioprospecting to attain information on species of fungus. <<http://pyrenomycetes.free.fr/xylariaceae/keydir/chemotaxonomy.htm>>;
- Screening for bioactive compounds from Fungi using PCR (Polymer Chain Reaction)-based data (Stadler and Hellwig 2005).
- In Australia, chemical prospecting for pharmaceuticals in molluscs is being studied as a tool for conservation (Benkendorff 1999) <<http://www.library.uow.edu.au/adt-NWU/public/adt-NWU20011204.154039/>>;
- The mining for biodiversity products can be carried out in conjunction with deep sea mining <<http://www.theworx.com/deepsea/mining.html>>;
- Herbal medicines <http://hcd2.bupa.co.uk/fact_sheets/html/herbal_medicine.html>.

Forensics

Primary species-occurrence data are a source of information for use in forensic research. Forensic science is based on protocols that require accurate identifications of organisms and precise distributional information (Page *et al.* 2004). Collections in natural history museums contain a massive store of DNA information that can be used in profiling and in determining locations, etc.

Gene Fragments

The identification of genetic fragments using DNA and comparing that with information held in museums or primary species databases is a key use in forensics.

Examples:

- Gene fragments were used to track rhino poachers by looking for genetic signatures in products such as powdered Asian medicines and Yemeni ornamental daggers. Fragments were able to identify not only the species, but individual game reserves that the horn came from. *New Scientist* 2411 (2003). <<http://www.newscientist.com/article.ns?id=mg17924110.700>>;
- Blood evidence from dogs has been used to convict murderers and rapists <http://www-ucdmag.ucdavis.edu/sp02/feature_2.html>;
- Analysis of DNA from an asthma inhaler was used to identify the administration of performance enhancing drugs to a race horse <http://www-ucdmag.ucdavis.edu/sp02/feature_2.html>;
- Forensic DNA sampling has established that an introduced colony of tamar wallabies living on Kawau Island, in New Zealand, is almost certainly comprised of the descendants of a wallaby subspecies that vanished from mainland South

Australia in the early 1900s. The subspecies is now being reintroduced to its original location

<http://www.bio.mq.edu.au/school/mag/intro/98bytes/may98/Bytes_May98.html>;

- DNA evidence is commonly used to convict for illegal trade in endangered species <<http://genetics.nbi.gov/forensics.html>>;
- DNA was used to identify contraband meat smuggled into the USA as red colobus monkeys (Nash 2001).

Plant material

The identification of plant material and the use of herbarium collections to identify them, is used in legal cases involving endangered species – plants that are the source of rugs, plants that help identify the scene of the crime, etc. Herbs and grasses on clothing can track the movement of criminals, or the origin of illegally transported objects, etc. Only by comparison with known material can definitive location and taxonomic information be determined.

Examples:

- By using a mass spectrometer to measure the ratios of carbon-12 to carbon-13, and nitrogen-14 to nitrogen-15, species of rhinoceros were identified in tracking poachers. The ratios vary depending on diet, and reveal whether horn came from white rhinos, which eat grasses, or black rhinos, which eat herbs and woody plants. Also by using optical emission spectrometers, the ratios of common trace elements such as iron and copper can identify locations where the material may have arisen *New Scientist* 2411 (2003). <<http://www.newscientist.com/article.ns?id=mg17924110.700>>;
- The identification of plant material, including cannabis is a common forensic use;
- The identification of leaves and fruits that may be found at a murder scene, or in a suspects car, etc. can help lead to a conviction <<http://www.sfu.ca/biology/faculty/mathewes/>>;
- The identification of plant parts in the intestinal tract of victims can aid in homicide investigations (Norris and Bock 2001);
- The identification of plant materials can be important in solving crimes (Lane *et al.* 1990).

Pollen

The pollen provides a key identification resource for use in forensic palynology. Forensic palynology is the study of pollen and powdered minerals. Their identification and location can be used to ascertain that a body or other object was in a certain place at a certain time.

Example:

- The Swedish Museum of Natural History maintains an international slide collection with more than 25000 pollen samples of different plant families <<http://www.nrm.se/pl/samling.html.en>>;
- The background and use of environmental profiling and forensic palynology (Wiltshire 2001) <http://www.bahid.org/docs/NCF_Env%20Prof.html>;
- The first conviction that used pollen analysis was in Austria in 1959. Pollen was used to identify a location where a body was buried using pollen in the mud on a suspect's boots <<http://www.saps.plantsci.cam.ac.uk/osmos/os23.htm>>;
- Pollen was used to identify the origin of a shipment of stolen Persian Rugs, although lack of suitable comparative species-occurrence data from Iran led to a failure to convict (Bryant and Mildenhall 2004) <<http://www.crimeandclues.com/pollen.htm>>.

Insects

Forensic entomology is used extensively to identify the time elapsed since the death of victims (Post-mortem Interval - PMI), to study whether bodies have been moved since death, to detect chemicals and poisons in bodies through the study of maggots, to track the movement of vehicles, and to determine the source of pest outbreaks (using houseflies and lesser houseflies) for city councils and health departments.

Examples:

- The use of forensic entomology <http://www.expertlaw.com/library/attyarticles/forensic_entomology.html>;
- Insects in legal investigations <<http://www.forensic-entomology.com/>>;
- The American Board of Forensic Entomology <<http://www.missouri.edu/~agwww/entomology/>>;
- Coleoptera and their significance in forensic entomology <<http://www.beetlelady.com/hister.html>>;
- Correct taxonomic identification of many insects and other Arthropoda can provide vital clues to the time and location of a death <http://www.bionet-intl.org/case_studies/case24.htm>
- The use of insects for determining Post-mortem Interval <http://www.absoluteastronomy.com/encyclopedia/F/Fo/Forensic_entomology.htm>;
- The use of maggots to determine time of death and to detect poisons and chemicals <<http://www.benecke.com/suntel.html>>.

Bird and Mammal Strikes

Bird strikes are a major problem with the safety of aircraft, etc. (Bird Strike Committee of the USA- <<http://www.birdstrike.org/events/signif.htm>>. The identification of these birds is essential to preventing future strikes, and species-occurrence data is an important tool in these identifications. Mammal strikes (e.g. large animal strikes with trains and road transport, etc.) can also be a problem in some areas.

Examples

- Bird Identification from the Smithsonian Institution (Dove et al. 2003). <<http://wildlife.pr.erau.edu/BirdIdentification.htm>>;
- Bird/Wildlife Strike Report Database, Environment Canada. <<http://www.tc.gc.ca/aviation/applications/birds/en/default.asp>>;
- Bird Strike Links from the International Bird Strike Committee <<http://www.int-birdstrike.com/links.html>>;
- German Bird Strike Committee (includes BIRDTAM). <<http://web.tiscali.it/birdstrike/>>;
- Bird Remains Identification System (BRIS) (Zoological Museum, Amsterdam) <<http://www.christ-media.de/cgi-bin/auswahl.cgi?basket=180757&wahl=2454629>>;
- Effect of bird strikes and the Bird Strike Information System (IBIS) <http://www.icao.int/icao/en/jr/5308_ar1.htm>.

Border Control and Wildlife Trade

Wildlife trade is a large industry, but one that invites illegal activities. Border control is used to prohibit the entry into countries of diseases, illegally traded wildlife such as endangered species, or products from endangered species such as ivory, pests such as might be transported unintentionally in wooden products, drugs, etc. Species-occurrence data are used

to provide border control agents with identification tools and means of identifying illegally traded and imported goods and to help them determine where they originated.

Border Controls and Customs

It is difficult for customs officers to know what is being illegally traded or not – what are pests that are prohibited etc. without good identification tools and access to primary species-occurrence data.

CITES

The Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) (<http://www.cites.org/>) aims to ensure that international trade in specimens of wild animals and plants does not threaten their survival. There are many listed species and groups of species, and it is difficult for customs officers to identify what might be an endangered species or not, and especially if a manufactured product or food has been derived from a CITES listed species.

Examples:

- Illegal bear parts seized by Customs Australia
<<http://forests.org/articles/reader.asp?linkid=32880>>;
- Federal agents target illegal bird trade
<<http://www.internationalparrotletsociety.org/smuggle.html>>;
- Illicit trade in orchids and wild plants (*Cites World* No 9, July 2002)
<<http://www.cites.org/eng/news/world/9.pdf>>;
- An illegal shipment of 9,300 live turtles was made in Hong Kong (*Traffic Bulletin* vol. 19 2002) <<http://www.traffic.org/bulletin/Nov2002/seizures3.html>>;
- CITES Identification tools and Guides
<http://www.cites.ec.gc.ca/eng/sct5/sct5_1_e.cfm>;
- Controlling the Shahtoosh trade in Tibet
<<http://www.met.police.uk/wildlife/new%20site%20docs/docs/shah.htm>>;

Illegal Fishing

Illegal fishing is of major concern to most maritime countries. Many of the species taken are CITES species, but others not.

Examples:

- Ecuador seizes illegal Galápagos Island shark fins
<<http://www.planetark.com/dailynewsstory.cfm/newsid/22244/newsDate/16-Sep-2003/story.htm>>;
- Illegal fishing threatens Galápagos Islands Waters
<http://news.nationalgeographic.com/news/2004/03/0312_040312_TVgalapagos.html>;
- Illegal fishing continues to grow (FAO)
<<http://www.fao.org/newsroom/en/focus/2004/47127/>>.

Drugs

Drugs and drug interception is another role for border control agents. The identification of drug and drug products is another use for primary species data.

Examples:

- Indian authorities have developed a database of Indian Medicinal Plants and species being traded as botanical drugs <<http://www.frlht-india.org/html/crg.htm>>;

- Regulating export of endangered medicinal plant species–Need for scientific rigour (Ved 1998) <<http://www.ias.ac.in/currsci/aug/articles8.htm>>.

Quarantine

Pests and Diseases

The importation of diseases and pests is of major interest and importance to agricultural industries as well as the general public. Again, the identification of pests and diseases can often be a problem for border control agents.

Examples:

- “Interception of potential agricultural, forest or medical pest species at U.S. borders will be greatly facilitated by access to a distributed network of taxonomic resources” (Page et al. 2004);
- Nematodes threaten US farmers <<http://www.hqusareur.army.mil/opm/aug04.htm>> and Mexican pecans <http://southwestfarmpress.com/mag/farming_nematodes_threaten_new/>;
- Please... don't bring pests or diseases with you to Australia <http://www.aust-immig-book.com.au/in_quarantine.html>;
- Australian Plant Pest Database <<http://appd.cmis.csiro.au/>>;
- In Namibia, identification of fruit flies allowed for more effective bilateral trade <http://www.bionet-intl.org/case_studies/case6.htm>.

Imported pets

The migration of people also means that pets are being transported across borders. Quarantine authorities need to monitor these for illegal importation, diseases, etc.

Wildlife Trade

Not all trade in wildlife is illegal, but controls of export and import permits requires knowledge and information on what species are being traded, and this requires primary species-occurrence data.

Examples:

- Wildlife trade and conservation in Australia <<http://www.deh.gov.au/biodiversity/trade-use/index.html>>;
- As part of its Wildlife Conservation Program, WWF Guianas is working with wildlife exporters and local governments to ensure that trade in wildlife is properly managed and based on the best scientific knowledge available. New Wildlife ID Manual <http://www.wwfguianas.org/Wildlife_IDman.htm>;
- Seahorses as wildlife trade – identification manuals <<http://www.worldwildlife.org/trade/seahorses.cfm>>;
- EU faces challenges in controlling Europe’s demand for wild animals and plants <http://www.traffic.org/news/enlarge_european.html>.

Education and Public Outreach

Education at all levels; along with public outreach are regular uses for primary species-occurrence data.

School level education

School level education at all levels benefits from integration with museums as well as being involved in school-level biodiversity data projects.

Examples:

- Museum School Partnership program (Doctoral Dissertation) (King 1998) <<http://home.iag.net/~ksking/muslearn.html>>;
- The GLOBE Program – hands on education and science program <http://www.globe.gov/globe_flash.html>;
- The Waterwatch program in Australia is a program conducted between museums, government, schools and the community to carry out biodiversity and habitat assessments of wetlands in their area <<http://www.waterwatch.org.au/>>;
- The Natural History Museum in London has an extensive education program – Exploring Biodiversity <<http://intern.nhm.ac.uk/eb/messages/probbrowser.shtml>>;
- In America, the National Zoo Biodiversity Monitoring Project works with school children to survey and monitor biodiversity in their area <<http://nationalzoo.si.edu/Publications/PressMaterials/BMPSchoolProjects.cfm>>;
- In Hungary, the Toad Action Group monitors amphibians with the aid of school children <http://www.virtualfoundation.org/publicboard/display.cgi?_Hungarian_amphibian_biodiversity_monitoring_EPCE_Hungary+archive>;
- In England, as part of the Stag Beetle Biodiversity Action Plan, schools were involved in recording and mapping the location of stag beetles across the country <http://www.lbp.org.uk/03action_pages/ac30_comms8.html>;
- Biodiversity for kids' teacher's kit <<http://www.bookshop.nsw.gov.au/pubdetails.jsp?publication=3403>>.

University level education

Universities are the training centers for the world's biodiversity specialists and most maintain museum and herbarium collections, and collect species-occurrence data as part of many of their courses.

Examples:

- Duke University hosts undergraduate students in a summer research program Bioinformatic and Phylogenetic Approaches to the Study of Plant and Fungal Biodiversity. <http://www.biology.duke.edu/reu/>
- The Xishuangbanna Tropical Botanical Garden in China runs a graduate student training courses in Asia in conjunction with a number of international universities <http://www.xtbg.ac.cn/english/PDF/gsxxtbg.pdf>

Training of Parataxonomists

The training of local peoples to be parataxonomists, requires extensive primary species data, including information on names, distributions, and often good image databases.

Examples:

- Training programs for parataxonomists have been developed by the National Biodiversity Institute (Inbio) in Costa Rica for use in the Guanacaste Conservation Area (Janzen *et al.* 1993, Janzen 1998);

- Training of local indigenous people has been carried out with insects in the Madang region of Papua New Guinea and in Guyana (Basset *et al.* 2000);
- In Hawaii parataxonomists are trained using insect processing by the Bishop Museum
<<http://www.bishopmuseum.org/research/natsci/guyana/LOGGING4.HTM>>;
- Brazilian Pollinators Initiative
<<http://www.mma.gov.br/port/sbf/chm/doc/pollinas.pdf>>;
- Taxonomic tools allow rapid problem solving by non-specialists
<http://www.bionet-intl.org/case_studies/case5.htm>.

Public awareness

The public are increasingly aware and becoming involved in their local environment (see also public participation programs below). Many organizations are now attempting to make it easier for people to find out about their natural environment and what is in it. This can be as simple as making available guidebooks so that people can identify the birds that visit their gardens, through to much more detailed environmental descriptions of their local areas.

Examples:

- The National Biodiversity Network (NBN) aims to make it easier for people to find out about their natural environment <<http://www.nbn.org.uk/>>;
- The North Australian Frogs Database System (Frogwatch) provides information to the public about frogs, cane toads and frog diseases to the Northern Territory public <<http://www.frogwatch.org.au/>>;
- The Biodiversity Conservation National Strategy and Action Plan of Uzbekistan aims to increase public awareness in biodiversity
<http://bpcsp-neca.brim.ac.cn/books/actpln_uzbek/>.

Books and materials

The publication of books and materials – local guide books to plants and animals, posters, screen savers and calendars all help in improving public awareness of biodiversity. Primary species-occurrence data are essential in helping develop these materials.

Examples:

- Australian mammals poster
<<http://www.bookshop.nsw.gov.au/pubdetails.jsp?publication=492>>;
- Fish Posters of the World <<http://www.fishposters.com/index.html>>;
- Animal Posters <<http://www.realtime.net/~raintree/gallery/posters.htm>>;
- Animal Screen Savers <<http://www.tnpsc.com/ssaver/animals.htm>>;
- NatureBase Screensavers of Western Australia
<<http://www.calm.wa.gov.au/screensavers/>>;
- Wildlife calendars from Africa
<<http://www.wildlife-pictures-online.com/wildlife-shopping-1.html>>;
- Lifemapper screensaver produces distribution maps
<<http://www.npaci.edu/online/v6.14/lifemapper.html>>.

Museum displays

Museum displays are a major source of education and public awareness. In recent years, museum displays have taken on the education role with increased vigour. Primary species-occurrence data play a key role in the development of these displays.

Examples:

- As early as 1995, the Field Museum saw benefits of automating collections records beyond scientific research. The integration of audio and textual data with visual images allows people to see exhibits from different museums and consider alternative interpretations from their homes or offices (Cohn 1995).
- North Carolina Nature Museums and Science Centers <<http://www.unc.edu/depts/cmse/museums.html>>;
- Australian Museum “What’s on” <<http://www.austmus.gov.au/visiting/whatson/>>.

Image Databases

Image databases are a valuable resource for development of virtual reference systems and on-line identification tools for biodiversity assessment (Oliver *et al.* 2000). For example, by linking to an on-screen reference system of insect specimen images, several parataxonomists working on the same taxon in remote laboratories can make identifications simultaneously, limiting the need for repeated handling and damage to valuable reference specimens (Oliver *et al.* 2000).

Examples:

- High-definition images are a core component of an on-line invertebrate identification network being established at the Macquarie University in Sydney (Oliver *et al.* 2000);
- Australian Plant Image Database <<http://www.anbg.gov.au/anbg/index-photo.html>>;
- Type Photos at the New York Botanic Gardens <<http://sciweb.nybg.org/science2/hcol/vasc/index.asp>>;
- Woodpecker images and sounds <http://www.infochembio.ethz.ch/links/en/zool_voegel_spechte.html>;
- Natural History image collections on the Web <<http://www.ucmp.berkeley.edu/collections/otherother.html>>;
- Digital Orthoptera Specimen Access (DORSA) <<http://www.dorsa.de/>>;
- Australasian Bird Image Database <<http://www.aviceda.org/abid/>>;
- Imagens da Biodiversidade Brasileira <<http://imagem.cria.org.br/>>;
- Digital Florilegium – part of the *New Endeavour* project <<http://www.invisible-consulting.com/endeavour/flora.php>>;
- Google Images <<http://www.google.com>>;

Public Participation Programs

Public participation conservation programs are becoming popular events. These can involve assistance in managing a river catchment area for conservation, water use and production; community planting of degraded areas; or conducting community-based conservation assessment.

Examples:

- The Calabash Program in Africa is a program to improve public participation in environmental assessments in Southern Africa <<http://www.sarpn.org.za/documents/d0000772/index.php>>;
- The Inter-American Environment Program of the Environment Law Institute is supporting and encouraging public participation in protecting landscapes in Argentina and in Conserving Community Lands in Mexico <<http://www2.eli.org/research/interamerican2.htm>>

- In Australia, the Federal Government funds community organizations to plant up areas of land degradation, high erosion, develop wildlife corridors, etc. Primary biodiversity data are used to identify suitable plant species and areas for planting <<http://www.landcareaustralia.com.au/>>;
- Again in Australia, Integrated Catchment Management Plans, sees community groups work closely with State and Federal Governments to plan and implement management plans for managing the resources, including water and biodiversity, and to balance that with agricultural production <<http://www.dlwc.nsw.gov.au/community/index.html>>;
- In Connecticut, in the United States, in the BioBlitz, program scientists work with community groups to carry out a rapid biodiversity assessment of local areas over intensive weekend programs (Lundmark 2003) <<http://www.mnh.uconn.edu/BioBlitz/>>;
- In the UK, the Natural History Museum the Walking with Woodlice project uses schools, local clubs, and individuals to survey UK woodlice <<http://www.nhm.ac.uk/interactive/woodlice/biodiversity.html>>;
- The Alcoa Frogwatch Program aims to involve a large number of people of all ages in actively helping to increase the quality of large-scale frog habitat <<http://frogs.org.au/frogwatch/>>;
- Total Catchment Management – public participation <<http://www.dlwc.nsw.gov.au/community/index.html>>;
- National Biodiversity Network’s Local Records Centers <<http://www.nbn-nfbr.org.uk/nfbr.php>>.

Tree of Life

The Tree of Life Web Project and similar collaborative projects provide information about the diversity of organisms on Earth, their history, and characteristics.

Examples:

- Tree of Life <<http://tolweb.org/tree/phylogeny.html>>;
- Diptera species pages <<http://www.diptera.org>>.

Ecotourism

Ecotourism is rapidly becoming one of the largest sources of income for many biodiversity-rich countries. UNEP recognises ecotourism as of special interest because of the role it can play in conservation, sustainability and maintenance of biological diversity (<http://www.uneptie.org/pc/tourism/ecotourism/home.htm>). Primary species-occurrence data are important in the development of good ecotourism programs – in the development of guidebooks, pamphlets, and information products and in helping countries determine suitable areas for use as ecotourism sites.

Valuing Ecotourism

One of the pressures against ecotourism is being able to put a value on biodiversity, conservation and ecotourism as an alternative to consumption and more intensive uses. But in many ecotourism projects, ecotourism and production can work side by side.

Examples:

- Valuing a tree for ecotourism
<<http://www.nuevomundotravel.com/nuevomundo.php?c=129>>;
- Valuing ecotourism in the Sierra Tarahumara region in Mexico
<http://www.srs.fs.usda.gov/econ/research/std43_8.htm>;
- Valuing Ecotourism as an Ecosystem Service (The Nature Conservancy)
<http://nature.org/event/wpc/files/drumm_presentation.pdf>;
- The Economics of “Eco Tourism”: A Galapagos Island Economy-wide Perspective (Taylor *et al.* 2002) <http://www.reap.ucdavis.edu/working_papers/jet-galapagos.pdf>.

Training Guides and Operators

The training of tour guides and tourism operators in understanding biodiversity is an area where primary species-occurrence data play a key role. Quite often, reference collections are made and kept at ranger stations, and carried by guides, and these require primary species data for identifications and training.

Examples:

- Eco-certified ecotourism in Australia
<http://www.ecotourism.org.au/eco_certification.asp>;
- Ecotourism certification workshops
<<http://www.planeta.com/ecotravel/tour/certification.html>>;
- Ecotourism Training Manual for Protected Area Managers (Strasdas 2002);
- Training Manual for Community-based Tourism (Inwent Zschortau, Leipzig, Germany) (Hausler and Strasdas 2003).

Guide Books

Guidebooks, pamphlets and other publications are an essential part of ecotourism and like guidebooks mentioned elsewhere are dependant upon species-occurrence data for their preparation.

Examples:

- Examples can be found in any bookshop or on the web of tour guidebooks, and most have an ecotourism section;
- A Guide to the Birds of Panama (Ridgely and Gwynne 1989);
- Costa Rica’s National Parks and Preserves: a visitor’s guide (Franke 1999).

Gardens, Zoos, Aquariums, Museums and Wildlife Parks

Botanical gardens, zoos, aquariums, wildlife parks and museums all play a part in ecotourism. Many new aquaria, for example include an underwater viewing area with access to the open sea. Most botanic gardens, zoos and wildlife parks maintain displays of the fauna and flora of the local regions and museums usually have extensive natural history displays. Most of these also have an education component. The labelling and information attached to these exhibits requires good data and information to prepare and maintain, including the names of the organisms involved and their distributions.

Examples:

- Monterey Bay Aquarium <<http://www.mbayaq.org/>>;

- Kirstenbosch National Botanical Garden <<http://www.nbi.ac.za/frames/kirstfram.htm>>;
- Jurong Bird Park, Singapore <<http://www.birdpark.com.sg/Main/>>;
- Jersey Zoo and Durrell Wildlife Conservation Trust <<http://www.durrellwildlife.org/>>;
- Smithsonian National Museum of Natural History <<http://www.mnh.si.edu/>>;
- Virtual Library: Museums around the world <<http://vlmp.museophile.org/world.html>>.

Art and History

Art has played an integral role in understanding and conservation of biodiversity. Most early scientific expeditions included an artist amongst their entourage to record the biodiversity. Today, artists continue to paint nature, and seek information on the names and locations of the subjects they paint. History is also a user of primary species-occurrence data. Early explorers were also natural historians and collected biodiversity specimens. With many centenaries and bicentenaries of these explorations coming up, many researchers are attempting to trace the steps of these early explorers and species-occurrence data are a major source of information for them.

History of Science —Tracking Explorers and Collectors

Early and modern explorers and scientists have deposited voucher specimens in natural history collections. “These specimens document the paths and objectives of the explorers and scientists over the centuries and provide a unique and irreplaceable source of historical data” (Page *et al.* 2004). As collections have aged, the year in which they were collected has become increasingly important (Winker 2004).

Examples:

- Nature's Investigator: The Diary of Robert Brown in Australia 1801-1805 (Vallance *et al.* 2001);
- Identifying collection patterns using Mexican bird specimens (Peterson *et al.* 1998);
- The *New Endeavour* is a project to revisit the landfalls of Captain James Cook's voyage in HMS Endeavour (1768-1771) <http://www.invisible-consulting.com/endeavour/>
- History of systematic botany in Australasia (Short 1990);
- Plant collectors in Brazil (Koch 2003) <http://splink.cria.org.br/collectors_db>;
- Lewis and Clarke Expedition in America <<http://www.cr.nps.gov/nr/travel/lewisandclark/encounters.htm>>;
- Australian Plant Collectors and Illustrators 1780s-1980s <<http://www.anbg.gov.au/bot-biog/index.html>>.

Art and Science

As mentioned above, art played an important part in early scientific discoveries. There were no cameras around and paintings were the only representation available of many plants and animals. Some early artist's interpretations of plants and animals was so detailed that many regard them as superior to many modern photographs.

- Sydney Parkinson was the artist on Cook's voyage of discovery to the South Seas in 1768-1771. He painted many animals (<http://pages.quicksilver.net.nz/jcr/~parkinson.html>), insects (<http://www.nhm.ac.uk/services/ibd/gfx/te/vod/17.jpg>) and plants (<http://internet.nhm.ac.uk/cgi-bin/perth/cook/>) and made the first sketch of a kangaroo http://www.nhm.ac.uk/library/art/drawingconclusions/more/hibiscus_more_info.htm#collection;
- Ferdinand Bauer (1760-1826) is regarded as one of the most remarkable botanical artists of all time (Bauer *et al.* 1976) <http://nokomis.com.au/html/biography.html>;
- John Gould's Birds of Asia <http://www.jadestonegallery.com/printgallery/gould/birdsofasia.htm>;
- Birds of the World – McClung Museum Special Exhibition 1997 <http://mcclungmuseum.utk.edu/spece/birds/birds.htm>;
- The World of Insects in Chinese Art: A Special Exhibition of Plant-and-Insect Paintings was an exhibition held at the National Palace Museum in Taiwan in 2001 <http://www.taiwanheadlines.gov.tw/20010816/20010814f2.html>.

Indigenous Art

Indigenous art and artefacts are a major source of income for indigenous peoples. Increasingly, artists want to supply information about the subjects of their art or the materials that make up the artefacts and products.

Examples:

- Untapped potential for cooperation between science and technology for mountain conservation in the Andes and Himalayas (Camino 2002) <http://www.mtnforum.org/resources/library/camia02a.htm>;
- *Canna indica* is commonly used in jewellery and for other purposes <http://waynesword.palomar.edu/pljune98.htm>;
- Nickernuts (*Caesalpinia bonduc*) are used for necklaces in Ecuador <http://waynesword.palomar.edu/nicker.htm>;
- Feathers have been a traditional adornment in many societies, The use of Birds-of-Paradise in Papua New Guinea is a good example (Frith and Beehler 1998)
- Yams are used for masks in Papua New Guinea <http://www.art-pacific.com/artifacts/nuginea/yamwoodo.htm>;
- Shells, feathers, grass twine and other materials are commonly used in indigenous art http://www.lostworldarts.com/new_page_2.htm;
- Wool is used in the Andes and Himalayas www.andeansoftware.com;
- Fibres are used in basket making <http://www.aotearoa.co.nz/flaxworks/>;
- Bamboo and other woods are used in making musical instruments http://www.canne-et-bambou.com/eng/bamboo_flutes.htm;
- Bark is used for paintings by indigenous Australians <http://www.aboriginalartonline.com/art/bark.html>.

Stamps

Most modern societies around the world use biodiversity on their stamps. These stamps often include scientific as well as common names, and stamp producers rely on primary species data for these identifications.

Examples:

- Australian Stamps: Bush Tucker
<http://www.auspost.com.au/philatelic/stamps/index.asp?link_id=2.608>;
- Birds on stamps <<http://www.birdtheme.org/regions/region.html>>;
- Kyrgyzstan animal stamps <<http://ecopage.freenet.kg/biodiversity/animals.html>>;
- Fijian stamps often include plants, insects and other animals
<http://www.stampsfiji.com/stamps/peregrine_falcon/index.html>.

Society and Politics

Many of the uses of species data in society and politics are covered under other topics; however, several uses do not seem to fit easily elsewhere.

Social Uses of Biodiversity

Biodiversity sits within the social context of human population – the competition between conservation and the need for food and shelter for survival is a never ending conflict. Many new studies are looking at the interaction between biodiversity and the social culture of humans.

Examples:

- Areas of high avian endemism also hold dense human populations and rapid rates of habitat loss, thus human population density and growth rates must be factored into conservation priority setting (Brooks 2001);
- Other important-and “often sensitive and contentious-parameters include the distributions of military conflict, refugee movements, timber and mining concessions, commodity production, bushmeat hunting, and the narcotics trade” (Brooks 2001);
- Several projects of the Biota/FAPESP Program in São Paulo, Brazil are examining social aspects of biodiversity
 - One study is looking at an environmental atlas to help in planning a balance between human activities and biodiversity
<<http://www.biota.org.br/projeto/index?show+192>>;
 - Another study is examining the use of natural resources for fishing, artefacts and for spiritual purposes by coastal inhabitants. The study is examining uses and local nomenclature as well as examining how the communities live and fish, and what effects their activities may have on the environment
<<http://www.biota.org.br/projeto/index?show+226>>;
- Mobilising European social research potential in support of biodiversity ecosystem management (SoBio) (European Centre for Nature Conservation)
<<http://www.ecnc.nl/doc/ecnc/press/070404.html>>;
- Unit for Social and Environmental Research – Chiang Mai University
<<http://www.sea-user.org/>>;

Anthropology and Language

Anthropological studies, and even some biological studies (Basset *et al.* 2000) have been attempting to link indigenous nomenclatural systems for species to the Linnaeus system.

Examples:

- In Papua New Guinea, studies are attempting to link local forest species nomenclature to species names as part of a project to train local people as parataxonomists and collectors of insects (Basset *et al.* 2000);
- Primary species data have been used to compare primate proteins <<http://www.bioquest.org/bioinformatics/module/tutorials/Anthropology/>>;
- Plant species data are used to identify species used in diets to track migration patterns (Newton-Fisher 1999) <http://www.budongo.org/nen1000/reprints/NewtonFisher_1999_diet.pdf>.

Ethnobiology

Local knowledge about useful plants and animals, extending back more than 300,000 years, is an important subject of research by ethnobotanists (Gómez-Pompa 2004) and ethnozoologists. The integration of this knowledge with distributional studies from primary species-occurrence data is an important area of research.

Examples:

- Some anthropology studies look at the use of plant and animal species in healing, medicine, and for food <<http://www.library.adelaide.edu.au/guide/soc/anthro/subj/med.html>>;
- Laboratory of Ethnobotany houses thousands of records of species used for food and medicine <<http://www.umma.lsa.umich.edu/ethnobotany/ethnobotany.html>>;
- Nuaulu Ethnozoology – A Systematic Inventory by Roy Allen from the University of Kent at Canterbury <http://lucy.ukc.ac.uk/csacpub/ellen_ch1.html>;
- *Acacia* in Australia: Ethnobotany and Potential Food Crop (Lister *et al.* 1996) <<http://www.hort.purdue.edu/newcrop/proceedings1996/v3-228.html>>;
- Ethnozoology of the Tsou People: Fishing with poison <http://tk.agron.ntu.edu.tw/Segawa1/fishing_poison.htm>
- Native Peoples, Plants and Animals <<http://www2.sfu.ca/halk-ethnobiology/>>;
- Ethnozoological Research on Reptiles on Mt. Kilimanjaro <<http://www.uni-bayreuth.de/departments/toek2/claudia/fEthnozoology.htm>>;
- Ethnobotany: Plants and People Interacting <<http://maya.ucr.edu/pril/ethnobotany/Start.html>>.

Data Repatriation

The Convention on Biological Diversity (CBD) calls for repatriation of information to countries of origin. More recently, the idea of one to one data repatriation of museum and herbarium collections has moved more toward the idea of data sharing, and especially through use of on-line data availability using portals such as the .

Examples:

- Report on study on data sharing with countries of origin (GBIF) <<http://www.gbif.org/Stories/STORY1079623109>>;
- Only 0.8% of the world's beetle researchers reside in Africa and few of the type specimens (Miller and Rogo 2001);
- The Natural History Museum in London is working in Chile on Access to Genetic Resources, Benefit Sharing and Traditional Knowledge <http://www.darwin.gov.uk/news/projects/access_gen.html>;
- The Natural History Museum is also working on the repatriation of herbarium data for the flora of Bahia, Brazil <<http://www.darwin.gov.uk/projects/details/7108.html>>

- Using Virtual museums to increase information repatriation and sharing *Whole Earth* 2000
<http://www.findarticles.com/p/articles/mi_m0GER/is_2000_Fall/ai_66240384>;
- A Mexican case study on a centralised database from World Natural History Collections (Navarro *et al.* 2003)
<http://journals.eecs.qub.ac.uk/codata/Journal/Contents/1_1/1_1pdf/DS105.Pdf>.

Biodiversity collecting

In many countries the development and expansion of protected areas is in some cases making it more difficult for scientists to collect and study biodiversity in these areas. Because of this, existing species-occurrence data will need to be relied on even more heavily in those areas where access for new collections may be restricted.

Recreational Activities

Recreational activities form another use for species-occurrence data. Many recreational activities involve biodiversity in one way or another – fishing, hunting, bird and whale watching, gardening, bushwalking, horse-riding, etc.

Recreational fishing

Recreational fishing is a large industry, and fishermen want to know what the fish is that they have caught – where certain species of fish occur and when, etc. All of this information is based on primary species-occurrence data.

Examples:

- Recreational fishers in Western Australia want habitat protected to improve recreational fishing <<http://www.recfishwest.org.au/PolicyFishHab.htm>>;
- In planning for zoning on the Great Barrier Reef, 36% of all submissions were from recreational fishers
<http://www.gbrmpa.gov.au/corp_site/management/zoning/rap/rap/overview/intro/re cfish.html>;
- Recreational fishing in Belarus is a major cause of biodiversity decline
<<http://www.iucn-ce.org.pl/documents/belarus.pdf>>;
- Recreational fishing is being considered in management of fishing resources in the Upper Paraná River Basin in Brazil
<<http://www.unep.org/bpsp/Fisheries/Fisheries%20Case%20Studies/AGOSTINHO.pdf>>.

Hunting

Like recreational fishers, hunters want to know what species they are hunting and where and when they occur. Conservations are also involved in knowing what species hunters are taking so that they can be taken into account in species management.

Examples:

- Hunting and Biodiversity in Atlantic Forest Fragments, São Paulo, Brazil
<<http://www.wildlifetrust.org/huntipe.htm>>;

- Extinction caused by hunting
<<http://www.virtualglobe.org/en/info/env/04/diversity07.html>>;
- Impacts of hunting on native species in New Zealand
<<http://www.biodiversity.govt.nz/picture/biodiversity/state/hunting.html>>;
- The North American Hunting Heritage Accord plans for sustainable hunting
<http://centralflyway.org/Hunting_Accord_Draft.html>.

Photography and Film-making

Photography of wildlife is another major recreational activity that relies on primary species-occurrence data for identification, and for determining where to find certain species to photograph, etc. Photographers are responsible for books, calendars, stamps, documentaries, etc. as well as on-line collections.

Examples:

- North American Nature Photography Association
<<http://www.nanpa.org/index.html>>;
- The Finnish Nature Photographers Association <<http://www.luontokuva.org/>>;
- The Discovery Channel <<http://dsc.discovery.com/>>;
- Nature and Wildlife Movies
<http://www.dropbears.com/b/broughsbooks/movies/nature_wildlife.htm>;
- David Attenborough Films
<http://www.bbc.co.uk/nature/programmes/who/david_attenborough.shtml>.

Gardening

Gardening is a passion among many, and the need to know what the plants are, essential to most gardeners. Books and magazines on gardening are constantly being marketed and all rely on species-occurrence data for their information. A number of people are also starting to get into organic gardening and are searching for species for growing.

Examples:

- Royal Horticultural Society
<<http://www.rhs.org.uk/research/biodiversity/index.asp>>;
- Gardening for Biodiversity <<http://www.english-nature.org.uk/news/story.asp?ID=257>>;
- Organic gardening books <http://supak.com/organic_gardening/organic.htm>;
- Australian Plants online – Society for Growing Australian Plants
<<http://farrer.riv.csu.edu.au/ASGAP/apoline.html>>.

Bushwalking, Hiking and Trekking

Bushwalking, hiking or trekking in natural areas is a common pastime that often involves people wanting to know what the species are that they pass.

Examples:

- Bushwalking in New South Wales
<<http://www.npansw.org.au/web/activities/bushwalking.htm>>;
- Hiking in Guatemala <http://www.guatemalaventures.com/hiking_tours.htm>;
- Hiking in Southeastern Arizona
<<http://www.geo.arizona.edu/geophysics/students/tinker/SEhiking.html>>;

- Trekking in Ecuador
<http://www.surtrek.com/ecuador/adventuretours/trek_podocarpus.htm>;
- Tramping in New Zealand <<http://www.enzed.com/tramp.html>>.

Bird Observing

Bird observing is a major recreational activity around the world, with many bird-observers clubs and bird activities. All rely on being able to identify the bird they have seen and thus rely on guide books and field guides created from primary species-occurrence data.

Examples:

- Birding.com <<http://www.birding.com/>>;
- National Audubon Society <<http://www.audubon.org/>>;
- Birding in Canada <<http://www.web-nat.com/bic/>>;
- Birds Australia <<http://www.birdsaustralia.com.au/>>;
- Birding Africa <<http://www.birding-africa.com/>>.

Human Infrastructure Planning

Planning of human infrastructure – roads, powerlines, subdivisions, etc. – requires species-occurrence data for finding the best place to build, and to do the least harm to the environment.

Risk Assessment

The building of roads and services requires risk assessment involving the most cost effective placement from both financial and ecological points of view. The management of weeds and hazardous vegetation on public lands, and the decision as to what species should be planted along roads and streets also involves risk assessment and species identifications.

Examples:

- Rights-of-Way Environmental Issues in Siting, Development and Management (Electric Power Research Institute)
<http://www.epri.com/destinations/descriptions/57_row.pdf>;
- Management of noxious weeds and hazardous vegetation on public lands – risk assessment for humans and non-target species
<<http://www.fs.fed.us/r3/projects/ro/ea-noxiousweeds/ea-weedsappa.html>>;
- Land use impact costs of transportation (Litman 1995)
<http://www.agenda21.ee/english/transport/landuse_costs_extern.pdf>;
- Significant costs in road maintenance can result from use of comprehensive biological survey <http://www.bionet-intl.org/case_studies/case19.htm>.

Landscaping

Tree roots of certain species can cause great damage to houses, sewage lines, etc. Street trees are often planted under powerlines and have to be trimmed at great cost as they get too tall, other species crack pavements and roads. Some species are more susceptible to damage in cyclones and tornadoes, etc. The selection of species that save energy and use less water can be important in some areas of the world. The identification of trees for planting in sensitive

locations and the identification of plants from their roots, etc. can require information from primary species-occurrence data.

Examples:

- Using dune vegetation to stop coastal dune erosion
<http://www.epa.qld.gov.au/environmental_management/coast_and_oceans/beaches_and_dunes/coastal_dunes/>;
- A Benefit-cost analysis of planting street tree species in Modesto, California (McPherson 2003) <<http://www.treelink.org/joa/2003/jan/01McPherson.pdf>>;
- Landscaping to save energy <<http://www.pioneerthinking.com/landscape.html>>;
- Tree roots a growing problem (South East Water Ltd, Melbourne, Australia) <<http://www.sewl.com.au/sewl/upload/document/treeroots.pdf>>;
- Windbreak trees for economic biodiversity (Stace 1995) <<http://www.newcrops.uq.edu.au/acotanc/papers/stace.htm>>;
- Planning tree windbreaks in Missouri <<http://muextension.missouri.edu/xplor/agguides/forestry/g05900.htm>>;
- Different species have differing abilities to ride out cyclones <http://www.plant.id.au/home/guide_view.aspx?id=15>.

Wild Animals and Infrastructure

Wild animals and human infrastructure always leads to clashes. Animals are killed on highways and roads, birds get sucked into aircraft engines, and wind turbines, dams stop species migrating up stream to spawn, etc. Primary species data are important in understanding species behaviour, locations etc.

Examples:

- Environment Canada is reducing wildlife roadkill <http://www.pc.gc.ca/pn-np/ab/banff/docs/routes/chap3/sec4/routes3d_e.asp>;
- The U.S Critter crossings reduce roadkills <<http://www.fhwa.dot.gov/environment/wildlifecrossings/index.htm>>;
- Avian interactions with utility structures, wind turbines and communication towers (EPRI's Destinations 2005) <<http://www.epri.com/destinations/product.aspx?id=309>>;
- Dams are being removed to save salmon <<http://www.wildsalmon.org/library/lib-detail.cfm?docID=300>>.

Building timbers

The selection of plant species for use in buildings for termite resistance, railway sleepers, bridges, fences, and power poles requires research into suitable species.

Examples:

- Termites and houses <<http://www.ces.ncsu.edu/depts/ent/notes/Urban/termites/termites.htm>>;
- Species of eucalypt are used in Australia for furniture, railway sleepers, bridge construction, flooring, etc. <http://www.tpcvic.org.au/page_timber_info.htm>;
- Acceptable species for use as power poles in Australia <<http://www.daleandmeyers.com.au/species.html>>;
- Incorrect identification of termites can be costly <http://www.bionet-intl.org/case_studies/case20.htm>.

Aquatic and Marine Biodiversity

Aquatic and marine biodiversity is largely covered under other topics above; however, they are covered separately here as there are some specific marine and aquatic biodiversity systems that require specific species-occurrence data.

Examples:

- Ocean Biogeographic Information System (OBIS)
<<http://www.coml.org/descrip/obis.htm>>;
- Gulf of Maine Biogeographic Information System Atlas (GMBIS)
<<http://gmbis.marinebiodiversity.ca/aconw95/aconscrip/gmbis.html>>;
- Riverine aquatic protected areas: protecting species, communities or ecosystem processes? (Koehn 2003);
- Census of Marine Life – “is a growing global network of researchers in more than 70 nations engaged in a ten-year initiative to assess and explain the diversity, distribution, and abundance of marine life in the oceans -- past, present, and future”
<<http://www.coml.org/coml.htm>>.

Conclusion

As seen throughout this document, uses for primary species-occurrence data are endless and touch just about every aspect of human endeavour, along with every part of the globe. They extend from uses for day to day survival such as food and shelter, through to education and learning, to pleasure and recreation. Most of us rely on these data without even thinking about them or even knowing they exist. But without them, whether held in museums or herbaria, in bird-observers databases or in survey databases held by Universities, individuals and corporations, we would not have the understanding of biodiversity that we have today, and our survival would be even further jeopardised than it already is.

We need to make maximum use of these data to better understand our biodiversity and our planet – to mitigate and monitor changes to our environment, to improve, conserve and sustainably use the resources we rely on and to educate and train future generations to appreciate and understand the biodiversity on which the data are based.

There are sure to be many uses that this document has missed, and it has been impossible to reference every example. It is hoped that the document may be made “live” in some format so that it can be kept updated and so that new uses can be added, possible by the on-line users of the data themselves.

References

- Akeroyd, J. and P. Wyse-Jackson (comps.). 1995. *A handbook for botanic gardens on the reintroduction of plants to the wild*. London: Botanic Gardens & Conservation International. 31 pp.
- Alcorn, J.B. (ed.). 1993. *Papua New Guinea Conservation Needs Assessment*. Washington: Conservation International.
- Almeida, A.C., Maestri, R., Landsberg, J.J., Scolforo, J.R.S., 2003. Linking process-based and empirical forest models in Eucalyptus plantation in Brazil in Amaro, A. and Tomé, M. (eds.), *Modelling Forest Systems*. CABI, Portugal, pp. 63-74.
- Amaral, A.C.Z. and Nallin, S.A.H. 2004. *Catálogo das espécies dos Annelida Polychaeta da Costa brasileira*. http://www.ib.unicamp.br/pesquisa/projetos/biota/bentos_marinho/7.htm. [Accessed 15 Apr. 2005].
- Amaro, A. and Soares, P. 2003. *Modelling Forest Systems*. CABI Publishing.
- Andrade, I., Morais, H.C., Diniz, I.R. and van den Berg, S. 1999. Richness and abundance of caterpillars on Byrsonima (Malpighiaceae) species in an area of cerrado vegetation in Central Brazil. *Rev. Biol. Trop. dic.* 47(4): 691-695. http://www.scielo.sa.cr/scielo.php?script=sci_arttext&pid=S0034-77441999000400005&lng=es&nrm=iso [Accessed 15 Apr. 2005].
- Asher, J., Warren, M., Fox, R., Harding, P., Jeffcoate, G. and Jeffcoate, S. 2001. *The Millenium Atlas of Butterflies in Britain and Ireland*. Oxford: Oxford University Press. <http://www.butterfly-conservation.org/index.html?/bnm/atlas/> [Accessed 15 Apr. 2005].
- Austin, M.P. 2002. Case Studies of the Use of Environmental Gradients in Vegetation and Fauna Modeling: Theory and Practice in Australia and New Zealand pp. 73-82 in Scott, M.J. et al. eds. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Washington: Island Press.
- Barrett, G., Silcocks, A., Barry, S., Cunningham, R. and Poulter, R. 2003. *The New Atlas of Australian Birds*. Melbourne, Australia, CSIRO Publishing. <http://birdsaustralia.com.au/atlas/> [Accessed 15 Apr. 2005].
- Basset, Y., Novotny, V., Miller, S.E. and Pyle, R. 2000. Quantifying Biodiversity: Experience with Parataxonomists and Digital Photography in Papua New Guinea and Guyana. *BioScience* 50(10): 899-908.
- Basson, M., Gordon, J.D.M., Large, P., Lorange, P., Pope, J and Rackham, B. 2002. The effects of fishing on deep-water fish species to the west of Britain. *JNCC Report* No 324, 150 pp.
- Bauer, F., Stearn, W.T. and Blunt, W. 1976. *Australian Flower Paintings of Ferdinand Bauer* London: Basilisk Press
- Belbin, L. 1993. Environmental representativeness, regional partitioning and reserve selection. *Biological Conservation* 66: 223-230.
- Belbin, L. 1994. *PATN: Pattern analysis package technical reference*. Canberra: CSIRO Division of Wildlife and Ecology.
- Benkendorff, K. 1999. *Bioactive molluscan resources and their conservation: Biological and chemical studies on the egg masses of marine molluscs*. Thesis, University of Wollongong <http://www.library.uow.edu.au/adt-NWU/public/adt-NWU20011204.154039/> [Accessed 15 Apr. 2005].
- Berenbaum, M.R. and Zangerl, A.R. 1998. Chemical phenotype matching between a plant and its insect herbivore. *Proc. Natl. Acad. Sci. USA* 95, 13743-13748. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24890> [Accessed 15 Apr. 2005].
- Berners-Lee, T. 1999. *Weaving the Web*. San Francisco, CA: Harper.
- Bickford, S.A., Laffan, S.W., de Kok, P.J. and Orthia, L.A. 2004. Spatial analysis of taxonomic and genetic patterns and their potential for understanding evolutionary histories. *J. Biogeogr.* 31: 1-23.
- BioCASE. 2003. *Biological Collection Access Service for Europe*. <http://www.biocase.org> [Accessed 12 Apr. 2005].
- Birds Australia. 2003. *Integrating Biodiversity into Regional Planning – The Wimmera Catchment Management Authority Pilot Project*. Canberra: Environment Australia. <http://www.deh.gov.au/biodiversity/publications/wimmera/methods.html>. [Accessed 15 Apr. 2005].
- Blakers, M., Davies, S.J.J.F. and Reilly, P.N. 1984. *The Atlas of Australian Birds*. Melbourne: Melbourne University Press.
- Booth, T.H. 1996. Matching Trees and Sites. Proceedings of an international workshop held in Bangkok, Thailand, 27-30 March 1995, *ACIAR Proceedings* No. 63.
- Boston, T and Stockwell, D. 1995. Interactive species distribution reporting, mapping and modelling using the World Wide Web. *Computer Networks and ISDN Systems* 28: 239-245.
- Bourque, D., Miron, G. and Landry, T. 2002. Predator-prey relationships between the nemertean *Cerebratulus lacteus* and the soft-shell clam, *Mya arenaria*: surface-exploration activity and qualitative observations on feeding behaviour. *Can. J. Zool.* 80(7): 1204-1211. [Accessed 19 Aug. 2004].

- Braby, M. 2000. *Butterflies of Australia. Their Identification, Biology and Distribution*. Melbourne: CSIRO Publishing.
- Breiman L. 1984. *Classification and regression trees*. Pacific Grove, CA: Wadsworth.
- Brooks, D.R. 2002. *Database for Inventory of Eukaryotic Parasites of Vertebrates of the Area de conservación Guanacaste, Costa Rica*. http://brooksweb.zoo.utoronto.ca/FMPro?-DB=CONTENT.fp5&-Format=intro.html&-Lay=Layout_1&-Error=err.html&content_id=1&-Find [Accessed 15 Apr. 2005].
- Brooks, D.R. and Hoberg, E.P. 2000. Triage for the biosphere: The need and rationale for taxonomic inventories and phylogenetic studies of parasites. *Comp. Parasitol.* 68: 1-25
- Brooks, T. 2001. Toward a blueprint for conservation in Africa. *BioScience* 51(8): 613-624.
- Bryant V.M. and Mildenhall, D.C. 2004. Forensic Palynology: A New Way To Catch Crooks. *Crimes and Clues. The Art and Science of Criminal Investigations*. <http://www.crimeandclues.com/pollen.htm>. [Accessed 15 Apr. 2005].
- Burrage, D.D., Branstetter, S.G., Graham, G. and Wallace, R.K. 1997. Development and Implementation of Fisheries Bycatch Monitoring Programs in the Gulf of Mexico. *Miss. Agric. Forest. Exper. Sta. Information Bulletin* 324: 103 pp. <http://www.rsca.org/docs/ib324.htm> [Accessed 15 Apr. 2005].
- Burton, H. 2001. Annual population estimates of Southern Elephant Seals at Macquarie Island from censuses made annually on October 15th., *Australian Antarctic Data Centre - SnowWhite Metadata* <http://www.aad.gov.au/default.asp?casid=3802> [Accessed 15 Apr. 2005].
- Busby, J.R. 1979. *Australian Biotaxonomic Information System. Introduction and Data Interchange Standards*. Canberra: Australian Biological Resources Study. 25pp.
- Busby, J.R. 1984. *Nothofagus cunninghamii* (Southern Beech) Vegetation in Australia. *Australian Flora and Fauna Series* No. 1. Canberra: Australian Biological Resources Study.
- Busby, J.R. 1991. BIOCLIM – a bioclimatic analysis and prediction system. pp. 4-68 in Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO.
- Butchart, S.H.M., Stattersfield, A.J., Bennun, L.A., Shutes, S.M., Akçakaya, H.R., Baillie, J.E.M., Stuart, S.N., Hilton-Taylor, C. and Mace, G.M. 2004. Measuring global trends in the status of biodiversity: Red List Indices for birds. *PLoS Biol* 2 (12): e383. <http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0020383>. [Accessed 14 Apr. 2005].
- Campbell, R.J. 1996. South American fruits deserving further attention. pp. 431-439. in Janick, J. (ed.) *Progress in new crops*. Arlington, VA: ASHS Press.
- Camino, A. 2002. An untapped potential for cooperation in science and technology for mountain conservation and sustainable development: the case of the Andes and the Himalayas *International Seminar on Mountains (ISM), Kathmandu, Nepal*. <http://www.mtnforum.org/resources/library/camia02a.htm> [Accessed 15 Apr. 2005].
- Cannell, M.G.R., Grace, J. and Booth, A. 1989. Possible impacts of climatic warming on trees and forests in the UK: a review. *Forestry* 62: 337-364.
- Carpenter, G., Gillison, A.N. and Winter, J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation* 2: 667-680.
- Catard, A.; Weimerskirch, H. 1998. Satellite tracking of petrels and albatrosses: from the tropics to Antarctica *Proceedings of the 22nd Ornithological Congress, Durban* 69(1-2): 152pp.
- CBD. 2004. *Alien Species*. Convention on Biological Diversity Secretariat. <http://www.biodiv.org/programmes/cross-cutting/alien/> [Accessed 15 Apr. 2005].
- CHAH 2002. *AVH - Australian's Virtual Herbarium*. Australia: Council of Heads of Australian Herbaria. <http://www.chah.gov.au/avh/avh.html> [Accessed 15 Apr. 2005].
- Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 in Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.
- Chapman, A.D. 2005a. *Principles of Data Quality*. Report for Global Biodiversity Information Facility, Copenhagen.
- Chapman, A.D. 2005b. *Principles and Methods of Data Cleaning*. Report for Global Biodiversity Information Facility, Copenhagen.
- Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 in Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.
- Chapman, A.D. and Milne, D.J. 1998. *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia
- Chapman, A.D., Bennett, S., Bossard, K., Rosling, T., Tranter, J. and Kaye, P. 2001. Environment Protection and Biodiversity Conservation Act, 1999 – Information System. *Proceedings of the 17th Annual Meeting of the Taxonomic*

- Databases Working Group, Sydney, Australia 9-11 November 2001*. Powerpoint:
http://www.tdwg.org/2001meet/ArthurChapman_files/frame.htm [Accessed 15 Apr. 2005].
- Chapman, A.D., Muñoz, M.E. de S. and Koch, I. (2005). Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context, *Biodiversity Informatics* 2: 24-41.
- Clements, M.A. and Ellyard, R.K. 1979. The symbiotic germination of Australian terrestrial orchids. *Amer. Orchid Soc. Bull.* 48: 810-815.
- CNLBSC. 2003. *Closure of Cod Fisheries – Action Plan*. Canada/Newfoundland and Labrador Business Service Centre, Government of Newfoundland and Labrador.
http://www.cbcs.org/nf/search/display.cfm?Code=6145&coll=NF_PROVBIS_E [Accessed 15 Apr. 2005].
- Coates, D. 1995. Inland capture fisheries and enhancement: status, constraints and prospects for food security. KC/FI/95/TECH/3. 82 p. *Contribution to the International Conference on the Sustainable Contribution of Fisheries to Food Security, Kyoto, Japan, 4-9 December 1995*, organized by the Government of Japan, in collaboration with the Food and Agriculture Organization of the United Nations (FAO).
- Cohn, J.P. 1995. Connecting by computer to collections. *BioScience* 45(8): 518-521.
- Colwell, R.K. 2000. *EstimateS. Statistical Estimation of Species Richness and Shared Species from Samples*.
<http://viceroy.eeb.uconn.edu/estimates> [Accessed 15 Apr. 2005].
- Colwell, R.K. and Lees, D.C. 2000. The mid-domain effect: geometric constraints on the geography of species richness. *TREE* 15:70–76.
- CONABIO. 2002. *Red Mundial de Información sobre Biodiversidad*. Mexico City: Comisión nacional para el conocimiento y uso de la biodiversidad. http://www.conabio.gob.mx/remib/doctos/remib_esp.html [Accessed 13 Apr. 2005].
- Conn, B.J. (ed.). 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data*. Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens. <http://www.bgbm.org/TDWG/acc/hispid30draft.doc> [Accessed 12 Apr. 2005].
- Conn, B.J. (ed.). 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data*. Version 4 – Internet only version. Sydney: Royal Botanic Gardens. <http://plantnet.rbgsyd.nsw.gov.au/Hispid4/> [Accessed 30 Jul. 2003].
- Coppack, T. and Both, C. 2003. Predicting life-cycle adaptation of migratory birds to global climate change. *Ardea* 90(3), special Issue: 367-378 <http://www.rug.nl/biologie/onderzoek/onderzoekgroepen/dierOecologie/publications/803Pdf.pdf> [Accessed 15 Apr. 2005]
- Costanza, R., Norton, B. and Haskell, B. (eds). 1992. *Ecosystem Health: New Goals for Environmental Management*. Island Press, Washington, D.C.
- Croft, J.R. (ed.). 1989. *HISPID – Herbarium Information Standards and Protocols for Interchange of Data*. Canberra: Australian National Botanic Gardens.
- CRIA. 2002. *speciesLink*. Campinas: Centro de Referência em Informação Ambiental. <http://splink.cria.org.br/> [Accessed 15 Apr. 2005].
- Crosbie, J.C. 1992. *Crosbie Announces First steps in Northern Cod Recovery Plan*. Press Release from Minister of Fisheries and Oceans, Canada. 1992. <http://www.stemnet.nf.ca/cod/announce.htm> [Accessed 15 Apr. 2005].
- Croxall, J.P., Briggs, D.R. and Prince, P.A. 1993. Movements and interactions of the Wandering Albatrosses: the roles of satellite tracking and direct observations *Sea Swallow* 42: 41-44
- Csuti, B., Polasky, S., Williams, P.H., Pressey, R.L., Camm, J.D., Kershaw, M., Kiester, A.R., Downs, B., Hamilton, R., Huso, M. and Sahr, K. 1997. A comparison of reserve selection algorithms using data on terrestrial vertebrates in Oregon. *Biological Conservation* 80: 83-97.
- Cunningham, D., Walsh, K. and Anderson, E. 2001. *Potential for Seed Gum Production from Cassia brewsteri*. RIRDC Project No. UCQ-12A. Kingston, ACT: Rural Industries Research and Development Corporation.
<http://www.rirdc.gov.au/reports/NPP/UCQ-12A.pdf>. [Accessed 15 Apr. 2005].
- Dallwitz, M.J. & T.A. Paine (1986). Users guide to the DELTA system. CSIRO Division of Entomology Report No. 13, pp. 3-6. *TDWG Standard*. (Periodic updates of this guide have been published.) <http://delta-intkey.com/>. [Accessed 14 Mar. 2005].
- Day, M.D. and Naser, S. 2000. Factors Influencing the Biological Control of *Lantana camara* in Australia and South Africa. *Proceedings of the X International Symposium on Biological Control of Weeds* 4-14 July 1999. Montana, USA.
<http://www.ppru.cornell.edu/weeds/Symposium/proceed/13pg897.pdf> [Accessed 18 Aug. 2004].
- Debach, P. 1974. Biological control by natural enemies. pp. 323. Vambridge: Cambridge University Press.
- DEH. 2000. *Environment Protection and Biodiversity Conservation (EPBC) Act 1999*. Canberra, Department of the Environment and Heritage. <http://www.deh.gov.au/epbc/index.html> [Accessed 15 Apr. 2005].

- DEH. 2004. *Threatened Ecological Communities*. Canberra: Department of the Environment and Heritage. <http://www.deh.gov.au/biodiversity/threatened/communities/index.html> [Accessed 15 Apr. 2005].
- Dexter, E.D., Chapman, A.D. and Busby, J.R. 1995. *The Impact of Global Warming on the Distribution of Threatened Vertebrates (ANZECC 1991)*. Report to Department of Environment Sport and Territories, Canberra. 163 pp
- Dimijiam, G.G. 1999. Pathogens and parasites: insights from evolutionary biology. *BUMC Proceedings* 12: 175-187. http://www.baylorhealth.edu/proceedings/12_3/12_3_dimijian.html [Accessed 15 Apr. 2005].
- Dove, C., Laybourne, R. Heacker-Skeans, M. 2003. Bird Identification. <http://wildlife.pr.erau.edu/BirdIdentification.htm> [Accessed 13 Apr. 2005].
- Duckworth, W.D., Genoways, H.H. and Rose, C.L. (1993). *Preserving Natural Science Collections: Chronicle of our Environment Heritage*. Washington, DC: National Institute for the Conservation of Cultural Property 140pp.
- Dunn, P.O. and Winkler, D.W. 1999. Climate change has affected the breeding date of tree swallows throughout North America. *Proc. R. Soc. London B. Biol. Sci.* 266(1437): 2487-2490
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?holding=np&cmd=Retrieve&db=PubMed&list_uids=10693819&dopt=Abstract [Accessed 15 Apr. 2005].
- Dynes, R.A. and Schlink, A.C. 2002. Livestock potential of Australian species of *Acacia*. *Conservation Science W. Aust.* 4(3): 117-124.
<http://science.calm.wa.gov.au/cswajournal/4-3/117-124.pdf> [Accessed 15 Apr. 2005].
- Edwards, J.L. 2004. Research and Societal Benefits of the Global Biodiversity Information Facility. *BioScience* 54(6): 485-486.
- Elkins, N., Reid, J. Brown, A. Robertson, D. and Smout, A.-M. 2003. The Fife Bird Atlas. Fife, UK. Fife Ornithological Atlas Group. http://www.the-soc.fsnet.co.uk/fife_bird_atlas.htm [Accessed 15 Apr. 2005].
- Englander, C. and Hoehn, P. 2004. *Checklist of Online Vegetation and Plant Distribution Maps*. Berkeley, CA: University of Berkeley Library <http://www.lib.berkeley.edu/EART/vegmaps.html> [Accessed 15 Apr. 2005].
- Faith, D.P. and Nicholls, A.O. 1996. *BioRap Volume 3: Tools for Assessing Biodiversity Priority Areas*. Canberra: The Australian BioRap Consortium.
- Faith, D.P. and Walker, P.A. 1996. Integrating conservation and development: effective trade-offs between biodiversity and cost in the selection of protected areas. *Biodiver. Conserv.* 5, 417-429.
- Faith, D.B. and Walker, P.A. 1997. Role of trade-offs in biodiversity conservation planning local management, regional planning and global conservation efforts. *Journal of Biosciences* 27(4): 393-407.
<http://www.ias.ac.in/jbiosci/jul2002/393.pdf> [Accessed 15 Apr. 2005].
- Faith, D.P., Walker, P.A., Ive, J., and Belbin, L. 1996. Integrating conservation and forestry production: Exploring trade-offs between biodiversity and production in regional land-use assessment. *Forest Ecology and Management* 85: 251-260.
- Faith, D.P., Walker, P.A., Margules, C.R., Stein, J. and Natera, G. 2001. Practical application of biodiversity surrogates and percentage targets for conservation in Papua New Guinea. *Pacific Conservation Biology* 6: 289-303
http://www.science.murdoch.edu.au/centres/others/pcb/toc/pcb_contents_v6.html [Accessed 15 Apr. 2005].
- Faith, D.P., Carter, G. Cassis, G. Ferrier, S. and Wilkie, L. 2003. Complementarity, biodiversity viability analysis, and policy-based algorithms for conservation. *Environmental Science and Policy* 6: 311-328.
http://www.amonline.net.au/systematics/pdf/faith_esap.pdf [Accessed 15 Apr. 2005].
- Falush, D. plus 17 other authors. 2003. Traces of human migrations in *Helicobacter pylori* populations. *Science* 299: 1582-1585
- Feller, A.E. and Hedges, S.B. 1998. Molecular Evidence for the Early History of Living Amphibians. *Molecular Phylogenetics and Evolution* 9(3): 509-516. <http://evo.bio.psu.edu/hedgeslab/Publications/PDF-files/101.pdf> [Accessed 15 Apr. 2005].
- Ferguson, N.M. and Anderson, R.M. 2002. Predicting evolutionary change in the influenza A virus. *Nat. Med.* 8(6): 562-3.
- Fjeldsa, J., Rahbek, C. 1997. Species richness and endemism in South American birds: Implications for the design of networks of nature reserves. pp. 466-482 in Laurence, W.L., Bierregaard, R. Jr., (eds) *Tropical Forest Remnants*. Chicago: Chicago University Press.
- Ferrier, S., Drielsma, M. Manion, G. and Watson, G. 2002. Extended statistical approaches to modelling spatial pattern in biodiversity in northeastern new South Wales. II. Community-level modelling. *Biodiversity and Conservation*. 11(12): 2309-2338.
- Fitzgerald, R.W. and Lees, B.G. 1992. The application of Neural Networks to the floristic classification of remote sensing and GIS data in complex terrain (I). *Proceedings 6th Australian Remote Sensing Conf., Wellington, N.Z.* V3; 2-10.
- Franke, J. 1999. *Costa Rica's National Parks and Preserves: a visitors guide*. Seattle, WA: The Mountaneers.

- Frith, C.B. and Beehler, B.M. 1998. *The Birds of Paradise*. Oxford, UK: Oxford University Press. 613pp.
- Gaston, K.J., Pressey, R.L. and Margules, C.R. 2002. Persistence and vulnerability: retaining biodiversity in the landscape and in protected areas. *J. Biosc.* (Suppl. 2) 27(4): 361-384.
- GBIF. 2004. *Data Portal*. Copenhagen: Global Biodiversity Information Facility. <http://www.gbif.net/portal/index.jsp>. [Accessed 15 Apr. 2005].
- Gillison, A.N. 2001. *Does biodiversity play a significant role in ecosystem function?* in Alternatives to Slash and Burn (ASB) Global Partnership ,Proceedings of Workshop Bringing the Landscape into Focus ,Developing a Conceptual Framework and Identifying Methods for ASB Work at the Landscape Scale Chiang Mai, Thailand <http://www.asb.cgiar.org/docs/SLUM%5C05-Ecological%20functions%20of%20biodiversity%5C05-2%20Does%20biodiversity%20play%20a%20significant.ppt> [Accessed 15 Apr. 2005].
- Gillison, A.N. and Carpenter, G. 1994. *A Generic Plant Functional Attribute Set and Grammar for Vegetation Description and Analysis*. Working Paper No. 3. Jakarta, Indonesia: CIFOR. http://www.cifor.cgiar.org/publications/pdf_files/WPapers/WP-03n.pdf [Accessed 15 Apr. 2005].
- Glasby, C.J. and Fauchald, K. 2003. *PoLiKEY. An information system for polychaete families and higher taxa version 2*. Canberra: ABRS. <http://www.deh.gov.au/biodiversity/abrs/online-resources/polikey/index.html> [Accessed 15 Apr. 2005].
- Gómez-Pompa. 2004. The Role of Biodiversity Scientists in a Troubled World. *BioScience* 54(3):217-225.
- Goodchild, M.F., Rhind, D.W. and Maguire, D.J. 1991. Introduction pp. 3-7 In: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
- Green, R.E. and Scharlemann, J.P.W. 2003. Egg and skin collections as a resource for long-term ecological studies. *Bull. British Ornithologists' Club*. 123A: 165-176 <http://www.boc-online.org/PDF/124GreenEggAndSkin.pdf> [Accessed 15 Apr. 2005].
- Güntsche, A. 2004. The ENHSIN Pilot Network – Implementation issues. Freie Universität Berlin, Botanischer Garten und Botanisches Museum Berlin-Dahlem. <http://www.bgbm.org/BioDivInf/projects/ENHSIN/PilotImplementation.htm> [Accessed 13 Apr. 2005].
- Harle, K.J., Hiodgson, D.A. and Tyler, P.A. 1999. Palynological evidence for Holocene palaeoenvironments from the lower Gordon River valley, in the World Heritage Area of southwest Tasmania. *The Holocene* 9(2): 149-162.
- Hambly, H. and Angura, T.O. 1996. Grassroot Indicators for Desertification. Experience and Perspectives from Eastern and Southern Africa. 180pp.
- Härmä, A. 2003. Automatic identification of bird species based on sinusoidal modelling of syllables *IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP'2003)*, Hong Kong. <http://www.acoustics.hut.fi/~sfagerlu/project/pubs/icassp03.pdf> [Accessed 15 Apr. 2005].
- Harris, W. 1994. Preliminary investigation of the suitability of *Cordyline australis* (Asphodeliaceae) as a crop for fructose production. *New Zealand Journal of Crop and Horticultural Science* 22: 439-451.
- Hastie, T.J. and Tibshirani, R.J. 1990. *Generalized Additive Models*, New York: Chapman and Hall
- Hausler, N and Strasdas, W. 2003. *Training Manual for Community-based Tourism*. Zschortau, Germany: Capacity-Building International.
- Hebert, P.D.N., Cywinska, A., Ball, S.L. and deWaard, J.R. 2003. Biological Identifications through DNA barcodes. *Proc. R. Soc. Lond.* B 270, 313-322.
- Higgins, D. and Taylor, W. 2000. *Bioinformatics: Sequence, Structure and Databanks – A Practical Approach*. Oxford University Press.
- Hijmans, R.J., Cameron, S., Para, J., Jones, P., Jarvis, A. and Richardson, K. (2004.). *Worldclim Version 1.2*. Berkeley, CA: Museum of Vertebrate Zoology. <http://biogeoberkeley.edu/worldclim/worldclim.htm> [Accessed 15 Apr. 2005].
- Hindmarsh, R. 2003. *Natural Resource Management Plan for the Brockman River Catchment*. Perth: Water and Rivers Commission. http://portal.environment.wa.gov.au/pls/portal/docs/PAGE/DOE_ADMIN/TECH_REPORTS_REPOSITORY/TAB1019581/WRM33.PDF [Accessed 15 Apr. 2005].
- Hnatiuk, R.J. 1990. Census of Australian Vascular Plants. *Australian Flora and Fauna Series* No. 11. Canberra: Australian Biological resources Study.
- Hoberg, E.P. 2002. Foundations for an integrative parasitology: collections archives and biodiversity informatics. *Comparative Parasitology* 69(2): 124-131.

- Hoffmeister, A.R., Fitzgerald, C.C., Ribot, E., Mayer, L.W. and Popovic, T. 2002. Molecular Subtyping of *Bacillus anthracis* and the 2001 Bioterrorism-Associated Anthrax Outbreak, United States. *Emerging Infectious Diseases* 8(10): 1111-1116. <http://www.cdc.gov/ncidod/EID/vol8no10/02-0394.htm> [Accessed 15 Apr. 2005].
- Howden, M., Hughes, L., Dunlop, M., Zethoven, I., Hilbert, D. and Chilcott, C. 2003. *Climate change impacts on biodiversity in Australia*. Canberra: CSIRO Sustainable Ecosystems. <http://www.deh.gov.au/biodiversity/publications/greenhouse/index.html> [Accessed 15 Apr. 2005].
- Hunt, K. 2001. Evolution and Mass Extinction in Freeman, S. and Herron, J.C. *Evolutionary Analysis*, 2nd edn. Prentice Hall.
- Iwu, M.M. 1996. Biodiversity prospecting in Nigeria: seeking equity and reciprocity in intellectual property rights through partnership arrangements and capacity building. *Journal of Ethnopharmacology* 51: 209-219.
- Janssen, D.L., Oosterhuis, J.E., Allen, J.L., Anderson, M.P., Kelts, D.G. and Wiemeyer, S.N. 1986. Lead poisoning in free-ranging California Condors. *J. Am. Vet. Med. Assoc.* 155: 1052-1056.
- Janzen, D.H. 1998. Gardenification of wildland nature and the human footprint. *Science* 279:1312-1313.
- Janzen, D.H. 2000. How to grow a wildland: the gardenification of nature pp. 521-529 in Raven, P.H. and Williams, T. (eds) *Nature and Human Society*. Washington, DC: National Academy Press.
- Janzen, D.H., Hallwachs, W., Jimenez, J., and Gamez R. 1993. The role of parataxonomists, inventory managers, and taxonomists in Costa Rica's national biodiversity inventory in Reid, V.W. et al. (eds). *Biodiversity Prospecting: Using Generic Resources for Sustainable Development*. Washington, DC: World Resources Institute.
- Kerry, K. 1999. *Satellite Tracking of Adelie Penguins Around Casey Station Antarctica*, Australian Antarctic Data Centre - SnoWhite Metadata <http://www.aad.gov.au/default.asp?casid=3802> [Accessed 15 Apr. 2005].
- Khan, Z.R., Ampong-Nyarko, K., Chiliswa, P., Hassanali, A., Kimani, S., Lwande, W., Overholt, W.A., Pickett, J.A., Smart, L.E., Wadhams, L.J. and Woodcock, C.M. 1997. Intercropping increases parasitism of pests. *Nature* 388: 631-632.
- King, K.S. 1998. *Museum School Partnership*. Doctoral Dissertation Indiana University. <http://home.iag.net/~ksking/muslearn.html> [Accessed 15 Apr. 2005].
- Koch, I. 2003. *Coletores de plantas brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 15 Apr. 2005].
- Koehn, J. 2003. Riverine aquatic protected areas: protecting species, communities or ecosystem processes? *Australian Society for Fish Biology*. http://www.asfb.org.au/research/mp/jk_aq_prot_areas.htm [Accessed 15 Apr. 2005].
- Komar, O., Robbins, M.B., Klenk, K., Blitvich, B.J., Marlenee, N.L., Burkhalter, K.L., Gubler, D.J., González, G., Peña, C.J., Peterson, A.T. and Komar, N. 2003. West Nile Virus Transmission in Resident Birds, Dominican Republic. *Emerging Infectious Diseases* 9(10): 1299-1302. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/Ketal_EID_2003.pdf [Accessed 13 Apr. 2005].
- Krishtalka, L. and Humphrey, P.S. 2000. Can Natural History Museums Capture the Future. *BioScience* 50(7): 611-617.
- Kristula, D. 2001. *The History of the Internet*. <http://www.davesite.com/webstation/net-history.shtml> [Accessed 13 Apr. 2005].
- Lane, M.A., Anderson, L.C., Barkley, T.M., Bock, J.H., Gifford, E.M., Hall, D.W., Norris, D.O., Rost, T.L. and Stern, W.L. 1990. Forensic Botany: Plants, perpetrators, pests, poisons and pot. *BioScience* 40: 34-39.
- Lee, J. 1997. *Floral Gems Coming to D.C.: Flower Power Saves South African Ecosystems*. USDA Agricultural Research Service News and Events <http://www.ars.usda.gov/is/pr/1997/971010.2.htm> [Accessed 15 Apr. 2005].
- Lawton, J.H., Bignell, D.E., Bolton, B., Bloemers, G.F., Eggleton, P., Hammond, P.M., Hodda, M., Holt, R.D., Larsen, T.B., Mawdsley, N.A., Stork, N.E., Srivastava, D.S. and Watt, A.D. 1998. Biodiversity indicators, indicator taxa and effects of habitat modification in tropical forest. *Nature* 391: 72-76 <http://invertebrates.ifas.ufl.edu/LawtonEtal.pdf> [Accessed 15 Apr. 2005].
- Lindenmayer, D.B. and Possingham, H.P. 1995. Modelling the impacts of wildfire on the metapopulation behaviour of the Australian arboreal marsupial, Leadbeater's possum, *Gymnobelideus leadbeateri*. *Forest Ecology and Management* 74:197-222
- Lindenmayer, D.B. and Possingham, H.P. 2001. *The risk of extinction: ranking management options for Leadbeater's Possum using population viability analysis*. Canberra: CRES, Australian National University 204 pp.
- Lindenmayer, D.B. and Taylor, M. 2001. *The Leadbeater's Possum Page*. Canberra: Australian National University <http://incres.anu.edu.au/possum/possum.html> [Accessed 15 Apr. 2005].
- Lister, P.R., Holford, P., Haigh, T. and Morrison, D.A. 1996. *Acacia* in Australia: Ethnobotany and potential food crop. p. 228-236 in Janick, J. (ed.), *Progress in new crops*. ASHS Press, Alexandria, VA <http://www.hort.purdue.edu/newcrop/proceedings1996/v3-228.html> [Accessed 15 Apr. 2005].

- Litman, T. 1995. Land use impact costs of transportation. *World Transport Policy and Practice* 1(4): 9-16. http://www.agenda21.ee/english/transport/landuse_costs_extern.pdf [Accessed 15 Apr. 2005].
- Lourie, S.A., Pritchard, J.C., Casey, S.P., Truong, S.K., Hall, H.J. and Vincent, C.J. 1999. The taxonomy of Vietnam's exploited seahorses (family Syngnathidae). *Biological J. Linn. Soc.* 66: 231-256. http://seahorse.fisheries.ubc.ca/pubs/Lourie_etal_vietnam.pdf [Accessed 13 Apr. 2005].
- Longmore, R. (ed.) (1986). Atlas of Elapid Snakes of Australia. *Australian Flora and Fauna Series No. 7*. Canberra: Australian Government Publishing Service.
- Lu, B.-R. 2004. Gene Flow from Cultivated Rice: Ecological Consequences. *IBS News Report* <http://www.isb.vt.edu/articles/may0402.htm> [Accessed 15 Apr. 2005].
- Lundmark, C. 2003. BioBlitz: Getting into Backyard Biodiversity. *BioScience* 53(4): 329.
- Lyne, A.M. 1993. *Leptospermum namadgiensis* (Myrtaceae), a new species from the Australian Capital Territory – New South Wales border area. *Telopea* 5(2): 319-324. <http://www.anbg.gov.au/projects/leptospermum/leptospermum-namadgiensis.html> [Accessed 15 Apr. 2005].
- Mackinnon, J. and De Wulf, R. 1994. Designing protected areas for giant pandas in China 127-142 in Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.
- Majer, J., Shattuck, S.O., Anderson A.N. and Beattie, A.J. 2004. Australian ant research: fabulous fauna, functional groups, pharmaceuticals, and the Fatherhood. *Australian Journal of Entomology* 43(3): 235
- MaNIS. 2001. *The Mammal Networked Information System*. <http://manisnet.org/manis> [Accessed 15 Apr. 2005].
- Margules, C.R. and Pressey, R.L. 2000. Systematic Conservation Planning. *Nature* **405**: 243-253.
- Margules, C.R., and Redhead, T.D. 1995. *BioRap: guidelines for using the biorap methodology and tools*. Canberra: CSIRO. 70pp.
- Margules, C.R., Nicholls, A.R. and Pressey, R.L. 1988. Selecting networks of reserves to maximise biodiversity. *Biological Conservation* 43: 63-76.
- Margules, C.R., Pressey, R.L. and Williams, P.H. 2002. Representing biodiversity: data and procedures for identifying priority areas for conservation. *J. Biosci.* 27(4): 309-326.
- Marshall, T.C., Sunnocks, P., Spalton, J.A., Greth, A. and Pemberton, J.M. 1999. Use of genetic data for conservation management: the case of the Arabian oryx. *Animal Conservation* 2: 269-278. <http://www.latrobe.edu.au/genetics/staff/sunnucks/homepage/papers/AnimalCons/Marshalletal98.pdf> [Accessed 15 Apr. 2005].
- McKenzie, G.M. and Busby, J.R. 1992. A quantitative estimate of Holocene climate using a bioclimatic profile of *Nothofagus cunninghamii* (Hook.) Oerst. *Journal of Biogeography* 19: 531-540.
- McKenzie, N.L. and Burbidge, A.L. 2002. *Australian Mammal Audit*. A Component of the National Land and Water Resources Biodiversity Audit.
- McPherson, E.G. 2003. A benefit-cost analysis of ten street tree species in Modesto, California, U.S. *Journal of Arboriculture* 29(1): 1-8
- Meisenheimer, P. 1998. *What is the Problem with Cod?* Guelph, ON: International Marine Mammal Association. 1998. <http://www.imma.org/codvideo/whatproblemcod.html> [Accessed: 15 Apr. 2005].
- Michelmore, F. 1994. Keeping elephants on the map: Case studies of the application of GIS for conservation pp. 107-123 in Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.
- Miller, S.E. 1991. Entomological collections in the United States and Canada: current status and growing needs. *American Entomologist* **37**(2): 77-84.
- Miller, S.E. and Rogo, L.M. 2001. Challenges and opportunities in understanding and utilisation of African species diversity. *Cimbebasia* 17: 197-218.
- Mills, J.N. and Childs, J.E. 1998. Ecologic Studies of Rodent Reservoirs: Their Relevance for Human Health. *Emerging Infectious Diseases* 4(4): 529-537. <http://www.cdc.gov/ncidod/eid/vol4no4/mills.htm> [Accessed 15 Apr. 2005].
- Mittermeier, R.A., Myers, N. and Mittermeier, C.G. 2000. *Hotspots: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions*. 430 pp. Chicago, IL: University of Chicago Press
- Monteiro, L.R. and Furness, R.W. 1998. Accelerated increase in mercury contamination in North Atlantic mesopelagic food chains as indicated by time series of seabird feathers. *Environmental Toxicology and Chemistry* 16(12): 2489-2493.
- Munro, R.K. and Williams, R.T. (eds). 1994. *Rabbit Haemorrhagic Disease: Issues for Biological Control*. Canberra: Bureau of Resource Sciences.

- Myers, N., Mittermeier, R.A., Mittermeier, C.G., Kent, J. and Fonseca, G.A.B. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403:853-858.
- Narosky, T. and Yzurieta, D. 2003. *Birds of Argentina and Uruguay. A Field Guide* 15th edn. Argentina: Vazquez Mazzini Editores.
- Nash, S. 2001. New Tools, Moon Tigers, and the Extinction Crisis. *BioScience* 51(9) 6.
- Nassar, N.M.A. 2003. Gene flow between cassava, *Manihot esculenta* Crantz, and wild relatives. *Genet. Mol. Res.* 2(4): 334-347.
http://www.funpecrp.com.br/gmr/year2003/vol4-2/gmr0047_full_text.htm [Accessed 15 Apr. 2005].
- Navarro-Sigüenza, A.G., Peterson, A.T. and Godillo-Martínez. 2003. Museums working together: The atlas of the birds of Mexico. *Bull. British Ornithologists' Club* 123A: 207-225
http://www.specifysoftware.org/Informatics/bios/biostownpeterson/NPG_BBOC_2003.pdf. [Accessed 15 Apr. 2005].
- Negrin, R.E.S., Moro, F.F., Alonso, G., Fernández, J.M.G. and Rodriguez, N.M.U. (eds). 2003. *Programa Nacional de Lucha Contra la Desertificación y la Sequía en la Republica de Cuba*. Havana, Cuba: Ministerio de Ciencia, Tecnología y Medio Ambiente. <http://www.unccd.int/actionprogrammes/lac/national/2003/cuba-spa.pdf> [Accessed 15 Apr. 2005].
- Neldner, V.J., Crossley, D.C. and Cofinas, M. 1995. Using Geographic Information Systems (GIS) to Determine the Adequacy of Sampling in Vegetation Surveys. *Biological Conservation* 73: 1-17
- Newton-Fisher, N.E. 1999. The diet of chimpanzees in the Budongo Forest reserve, Uganda. *Afr. J. Ecol.* 37: 344-354
http://www.budongo.org/nen1000/reprints/NewtonFisher_1999_diet.pdf [Accessed 15 Apr. 2005].
- Nghia, N.H. and Kha, L.D. 1996. Acacia species and provenance selection for large-scale planting in Vietnam. *Proceedings of 1996 QFRI – IUFRO Conference, Tree Improvement for Sustainable Tropical Forestry*, Caloundra, Queensland. <http://www.forests.qld.gov.au/resadv/research/qfri6.htm> [Abstract accessed 15 Apr. 2005].
- Nicholls, N. 1997. Increased Australian wheat yield due to recent climate trends. *Nature* 387: 484-485.
- Nicholls, O.W., Provan, D.J.M., Cole, M.M. and Tooms, J.S. 1965. Geobotany and geochemistry in mineral exploration in the Dugald River Area, Cloncurry District, Australia. *Trans. Inst. Mining and Metallurgy* 74: 695-799.
- NISO. 2002. *Z39.50 Resource Page*. Bethesda, MD: National Information Standards Organization.
<http://www.niso.org/z39.50/z3950.html> [Accessed 15 Apr. 2005].
- Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes in Longmore, R.C. (ed). *Atlas of Australian elapid snakes. Australian Flora and Fauna Series No. 7*: 4-15. Canberra: Australian Government Publishing Service.
- Nix, H.A. and Switzer, M. (eds). 1991. *Rainforest Animals: Atlas of Vertebrates Endemic to Australia's Wet Tropics. Kowari 1* Canberra: Australian National Parks and Wildlife Service.
- Nix, H.A.; Faith, D.P.; Hutchinson, M.F.; et al. 2000. *The BioRap Toolbox: A National Study of Diversity Assessment and Planning for Papua New Guinea*. Canberra: CRES, Australian National University.
- Norris, D.O. and Bock, J.H. 2001. Method for examination of fecal material from a crime scene using plant fragments. *Journal of Forensic Investigation* 51: 367-377.
- NRC. 2003. *Countering Agricultural Bioterrorism*. National Research Council (NRC). Washington, DC: National Academy Press.
- NSW National Parks and Wildlife Service. 2003. *Draft NSW and National Recovery Plan for the Tarengo Leek Orchid (Prasophyllum petilum)* Hurstville, NSW: NSW National Park and Wildlife Service.
- Ntiamoa-Baidu, Y. 1997. Wildlife and food security in Africa. *FAO Conservation Gude* 33.
<http://www.fao.org/docrep/W7540E/w7540e00.htm> [Accessed 15 Apr. 2005].
- Odhiambo, T.R. 1977. Entomology and the problems of the tropical world (pp. 52-59) in *Proceedings of the XV International Congress of Entomology*. College Park, Maryland: Entomological Society of America. 824 pp.
- OECD. 1999. *Final Report of the Megascience Forum Working Group on Biological Informatics*. Paris: OECD.
- Olesen, J.E. 2001. Climate Change and Agriculture in Denmark in Jørgensen, A.M.K.; Fenger, J.; Halsnæs, K. (eds), *Danish contributions*. Copenhagen: Danish Meteorological Institute pp. 191-206 <http://glwww.dmi.dk/f+u/publikation/dkc-publ/klimabog/CCR-chap-12.pdf> [Accessed 15 Apr. 2005].
- Oliver, I., Pik, A., Britton, D. Dangerfield, M.J., Colwell, R.K. and Beattie, A.J. 2000. Virtual Biodiversity Assessment Systems. *BioScience* 50(5): 441-450.
- Omasa, K., Saji, H., Youssefian, S. and Kondo, N. 2002. *Air Pollution and Plant Biotechnology – Prospects for Phytomonitoring and Phytoremediation*. Springer Verlag 455 pp.

- Page, L., Funk, V., Jeffords, M., Lipscomb, D., Mares, M. and Prather, A. (eds). 2004. Workshop to Produce a Decadal Vision for Taxonomy and Natural History Collections. *Report to the U.S. National Science Foundation Biodiversity Surveys and Inventories Program*. http://www.flmnh.ufl.edu/taxonomy_workshop/NSF_Workshop_Report_3-08-04.pdf [Accessed 15 Apr. 2005].
- Panetta, F.D. and Mitchell, N.D. 1991. Bioclimatic prediction of the potential distribution of some weeds prohibited entry to New Zealand *N.Z. J. Agric. Res.* **34**: 341-350.
- Parmesan, C., Rurholm, N., Stefanescu, C., Hill, J.K., Thomas, C.D., Descimon, H., Huntley, B., Kaila, L., Kullberg, J., Tammaru, T., Tennent, W.J., Thomas, J.A. and Warren, M. 1999. Poleward shift of butterfly species' ranges associated with regional warming. *Nature* 399: 579-583 http://www.biosci.utexas.edu/IB/faculty/parmesan/pubs/Parm_Ntr_99.pdf [Accessed 15 Apr. 2005].
- Pereira, R.S. 2002. *Desktop Garp*. Lawrence, Kansas: University of Kansas Center for Research. <http://beta.lifemapper.org/desktopgarp/> [Accessed 15 Apr. 2005].
- Pergams, O.R.W. and Nyberg, D. 2001. Museum collections of mammals corroborate the exceptional decline of prairie habitat in the Chicago region. *Journal of Mammalogy* 82(4): 984-992 <http://home.comcast.net/~oliver.pergams/ratio.pdf> [Accessed 15 Apr. 2005].
- Perkins L. and Swayne, D. 2002. Susceptibility of Laughing Gulls (*Larus atricilla*) to a H5n1 and a H5n3 Highly Pathogenic Avian Influenza Virus. *Avian Diseases* 46(4): 877-885.
- Perring, F.H. and Walters, S.M. eds. 1962. *Atlas of the British Flora*, London: Nelson - for Botanical Society of the British Isles
- Peters, D. and Thackway, R. 1998. *A New Biogeographic Regionalisation for Tasmania*. Hobart: Parks and Wildlife Service <http://www.gisparks.tas.gov.au/dp/newibra/Title&Background.htm> [Accessed 15 Apr. 2005].
- Peterson, A.T. 2003. Predicting the Biogeography of species Invasions via Ecological Niche Modeling. *Quarterly Rev. Biol.* 78(4): 419-433. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/P_QRB_2003.pdf [Accessed 13 Apr. 2005].
- Peterson, A.T., Navarro-Sigüenza, A.G. and Benitez-Diaz, H. 1998. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis*, 140:288-294.
- Peterson, A.T., and Vieglais, D.A. 2001. Predicting species invasions using ecological niche modeling. *BioScience* **51**: 363-371 http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PV_B_2001.pdf. [Accessed 15 Apr. 2005].
- Peterson, A.T., Ortega-Huerta, M.A., Bartley, J., Sánchez-Cordero, V., Soberón, J., Buddemeier, R.H. and Stockwell, D.R.B. 2002a. Future projections for Mexican faunas under global climate change scenarios. *Nature* **416**: 626-629. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/Petal_N_2002.pdf. [Accessed 15 Apr. 2005].
- Peterson, A.T., Ball, L.G. and Cohoon, K.P. 2002b. Predicting distributions of Mexican birds using niche modelling methods. *Ibis* 144: e27-e32. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/NPG_BBOC_2003.pdf [Accessed 15 Apr. 2005].
- Peterson, A.T., Scachetti-Pereira, R. and Kluza, D.A. 2003a. Assessment of Invasive Potential of *Homalodisca coagulata* in Western North America and South America. *Biota Neotropica* 3(1). <http://www.biotaneotropica.org.br/v3n1/en/download?article+BN00703012003+item> [Accessed 15 Apr. 2005].
- Peterson, A.T., Vieglais, D.A. and Andreassen, J.. 2003b. Migratory birds as critical transport vectors for West Nile Virus in North America. *Vector Borne and Zoonotic Diseases*, 3:39-50. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PVA_VBZD_2003.pdf [Accessed 15 Apr. 2005].
- Peterson, A.T., Scachetti-Pereira, R. and Hargrove, W.W. 2004. Potential geographic distribution of *Anoplophora glabripennis* (Coleoptera: Cerambycidae) in North America. *American Midland Naturalist* 151: 170-178. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PSH_AMN_2004.pdf [Accessed 15 Apr. 2005].
- Pettitt, C. 1991. What Price Natural History Collections, or 'Why de we need all these bloody mice?' *Mus. Journal* 91(8): 25-28. <http://fenscore.man.ac.uk/Uses/cwpmusjap.htm> [Accessed 15 Apr. 2005].
- Pimentel, D. (ed.). 2002. *Biological Invasions: Economic and Environmental Costs of Alien Plant, Animal, and Microbe Species*. Boca Raton, FL: CRC Press.
- Pimentel, D.; Lach, L.; Zunigar, R.; and Morrison, D. 1999. *Environmental and Economic Costs Associated with non-Indigenous Species in the United States*. Ithaca, NY: College of Agricultural and Life Sciences, Cornell University. 1999. http://www.news.cornell.edu/releases/Jan99/species_costs.html [Accessed 15 Apr. 2005].
- Pimentel, D.; Lach, L.; Zunigar, R.; and Morrison, D. 2000. Environmental and economic costs of nonindigenous species in the United States. *BioScience* 50(1): 53-65.
- Platt, T.R. 2000. *Neopolystoma fentoni* n. sp. (Monogenea: Polystomatidae) a parasite of the conjunctival sac of freshwater turtles in Costa Rica. *Memorias do Instituto Oswaldo Cruz* 95: 833-837. <http://brooksweb.zoo.utoronto.ca/pdf/Neopolystoma%20fentoni.pdf>. [Accessed 15 Apr. 2005].

- Pouliquen-Young, O. and Newman, P. (1999). *The Implications of Climate Change for Land-Based Nature Conservation Strategies*. Final Report 96/1306, Australian Greenhouse Office, Environment Australia, Canberra, and Institute for Sustainability and Technology Policy, Murdoch University, Perth, Australia, 91 pp.
- Pujari, G.N. and Shrivastava, J.P. 2001. High bioassay values in *Terminalia alata* leaves: indication of Cu mineralisation in Malanjkhand Granitoid, Central India. *Chemical Speciation and Bioavailability* 13(4): 97-111.
- Purvis, A., Gittleman, J.L., Cowlshaw, G. and Mace G.M. 2000. Predicting extinction risk in declining species. *Proc. Roy. Soc. Lond. B* 267: 1947-1952.
- Queensland Museum. 2004. *Saving Lives: Queensland Museum Collections*.
<http://www.qmuseum.qld.gov.au/features/snakes/saving.asp> [Accessed 15 Apr. 2005].
- Raina, S.K. (ed). 2000. *The economics of apiculture and sericulture modules for income generation in Africa*. Nairobi: ICIPE Science Press. 86 pp.
- Ratcliffe, D.A. 1967. Decrease in eggshell weight in certain birds of prey. *Nature*. 215: 208-210.
- Raxworthy, C.J., Martinez-Meyer, E., Horning, N., Nussbaum, R.A., Schneider, G.E., Oregua-Huerta, M.A. and Peterson, A.T. 2003. Predicting distributions of known and unknown reptile species in Madagascar. *Nature*. 426: 837-841.
- Redhead, T., Mummery, J. and Kenchington, R. (eds). 1994. *Options for a National Program on Long-Term Monitoring of Australian Biodiversity*. Canberra: CSIRO & Department of Environment, Sport and Territories.
- Ridgely, R.S. and Gwynne, J.A. 1989. *A Guide to the Birds of Panama* 2nd edn. Princeton, NJ: Princeton University Press.
- Rivas, M., Warner, J., Bermúdez, M. 1998. Presencia de micorrizas en orquídeas un jardín botánico neotropical *Rev. biol. Trop.* 46(2): http://www.scielo.sa.cr/scielo.php?pid=S0034-77441998000200004&script=sci_arttext&tlng=es
- Robinson, T.P., Rogers, D.J. and Williams, B.G. 1997. Mapping tsetse habitat suitability in the common fly belt of southern Africa using multivariate analysis of climate and remotely sensed vegetation data. *Medical and Veterinary Entomology* 11, 235-245.
- Rodgers, J. A. 1990. Breeding chronology and clutch information for the wood stork from museum collections. *J. Field Orn.* 61: 47-53.
- Schell, D. 2000. Declining carrying capacity in the Bering Sea: Isotopic evidence from whale baleen. *Limnology and Oceanography* 43: 459-462.
- Schmitt, C.J. and Bunck, C.M. 1995. *Persistent Environmental Contaminants in Fish and Wildlife*. USGS.
<http://biology.usgs.gov/s+t/noframe/u208.htm> [Accessed 14 Aug. 2004].
- Sekhran, N. and Miller, S. 1995. *Papua New Guinea Country Study on Biological Diversity*. Port Moresby: Department of Environment and Conservation.
- Shalk, P.H. and Heijman, P. 1996. ETI's Taxonomic *Linnaeus II* Software. A New Tool for Interactive Education. *Uniserve-Science News* 3: <http://science.uniserve.edu.au/newsletter/vol3/schalk.html> [Accessed 15 Apr. 2005].
- Shapiro, B. and Cooper, A. 2003. Beringia as an Ice Age genetic museum. *Quaternary Research* 59: 94-100.
- Shiembo, P.N. 2002. The sustainability of Eru (*Gnetum africanum* and *Gnetum buchholzianum*) an exploited non-wood forest production from the forests of Central Africa in *RATTAN Current research Issues and Prospects for Conservation and Development*. FAO. <http://www.fao.org/docrep/X2161E/x2161e06.htm>. [Accessed 15 Apr. 2005].
- Schulze, R., Meigh, J. and Horan, M. 2001. Present and potential future vulnerability of eastern and southern Africa's hydrology and water resources. *South African Journal of Science* 97: 150-160.
- Short, P.S. (ed.). 1990. *History of systematic botany in Australia*. Proceedings of a symposium held at the University of Melbourne 25-27 May 1988. South Yarra, Vic.: Australian Systematic Botany Society. 326 pp.
- Siddall, M.E. 1997. The AID Pandemic is New, but is HIV **Not** New? *Cladistics* 13: 266-273.
<http://research.amnh.org/~siddall/HIV.pdf> [Accessed 15 Apr. 2005].
- Siqueira, M.F. de, and Peterson, A.T. 2003. Global climate change consequences for cerrado tree species. *Biota Neotropica*, 3(2): <http://www.biotaneotropica.org.br/v3n2/en/download?article+BN00803022003+item> [Accessed 15 Apr. 2005].
- Soberón, J. 2004. *The National Biodiversity Information System of Mexico*
http://circa.gbif.net/Public/irc/gbif/pr/library?!=power_point/presentations_assembly/4_soberon_pps/ [Accessed 23 Aug 2004].
- Soberón, J., Golubov, J. and Sarakhán, J. 2000. Predicting the Effects of *Cactoblastis cactorum* Berg on the *Platyopuntia* of Mexico: A Model on the Route of Invasion pp. 95-97 in *Assessment and Management of Alien Species that Threaten Ecosystems, Habitats and Species*. CBD Technical Series No. 1. Montreal, Canada: Convention on Biological Diversity. . Copenhagen: GBIF. Powerpoint presentation (17 Mb) <http://www.biodiv.org/doc/publications/cbd-ts-01.pdf> [Accessed 15 Apr. 2005].

- Soberón, J., Golubov, J. and Sarakhán, J. 2001. The Importance of *Opuntia* in Mexico and Routes of Invasion and impact of *Cactoblastus cactorum* (Lepidoptera: Pyralidae). *Florida Entomologist* 84(4): 486-492.
- Soberón, J. Huerta-Ocampo, E., Arriaga-Cabrera, L. 2002. The Use of Biological Databases to Assess the Risk of Gene Flow: The Case of Mexico in *LMOS and the Environment*, OECD. <http://www.oecd.org/dataoecd/40/56/31526579.pdf> [Accessed 15 Apr. 2005].
- Sondström, O., Larsson, Å., Andersson, J., Appelberg, M., Bignert, A., Ek, H., Förlin, L. and Olsson, M. 2004. *Integrated fish monitoring in Sweden*. Helsinki: Helcom Monas Coastal Fish Monitoring [http://www.helcom.fi/dps/docs/documents/Monitoring%20and%20Assessment%20Group%20\(MONAS\)/MONAS%20Coastal%20Fish%20Monitoring%201.%202004/3-4.pdf](http://www.helcom.fi/dps/docs/documents/Monitoring%20and%20Assessment%20Group%20(MONAS)/MONAS%20Coastal%20Fish%20Monitoring%201.%202004/3-4.pdf) [Accessed 15 Apr. 2005].
- Sonnekus, I.P. and Breytenbach, G.J. 2001. Conservation business: sustaining Africa's future. *Koedoe* 44: 105-123.
- SourceForge. 2004. *Distributed Generic Information Retrieval (DiGIR)*. <http://digir.sourceforge.net/> [Accessed 15 Apr. 2005].
- Southwell, C., Meyer, L. 2003. The utility of satellite remote sensing for identifying the location and size of penguin breeding sites in Antarctica: a review of previous work and specifications of some current satellite sensors. *CCAMLR Scientific Abstracts*. WG-EMM-03/51 19 http://cs-db.aad.gov.au/proms/public/report_project_public.cfm?project_no=2205 [Accessed 15 Apr. 2005].
- Stace, P. 1995. Winbreak trees for economic biodiversity: a habitat for pests, predators and crop pollinators ACOTANC – 95. *The Sixth Conference of the Australian Council on Tree and Nut Species, Lismore, Australia* <http://www.newcrops.uq.edu.au/acotanc/papers/stace.htm> [Accessed 15 Apr. 2005].
- Stadler, J., Mungai, G. and Brandl, R. 1998. Weed invasion in East Africa: insights from herbarium records. *African Journal of Ecology* 36: 15-22.
- Stadler, M. and Hellwig, V. 2005. PCR-based Data and Secondary Metabolites as Chemotaxonomic Markers in High-Throughput Screening for Bioactive Compounds from Fungi. In *Handbook of Industrial Mycology* (Z. An, ed.) New York: Marcel Dekker. 269pp.
- Stattersfield, A.J., Crosby, M.J., Long, A.J. and Wege, D.C. 1998. *Endemic bird areas of the world: priorities for biodiversity conservation*. Birdlife International.
- Stockwell, D. and Peters, D. 1999. "The GARP modelling system: problems and solutions to automated spatial prediction." *International Journal of Geographical Information Science* 13(2): 143-158.
- Strasdas, W. 2002. *The Ecotourism Training Manual for Protected Area Managers*. Zschortau, Germany: German Foundation for International Development.
- Suarez, A.V., Holway, D.A. and Case, T.J. 2001. Patterns and spread in biological invasions dominated by long-distance jump dispersal: Insights from Argentine ants. *Proceedings of the National Academy of Sciences* 98: 1095-1100.
- Suarez, A.V. and Tsutsui, N.D. 2004. The Value of Museum Collections for Research and Society. *BioScience* 54(1): 66-74.
- Swaney, D. 1999. *Zimbabwe, Botswana & Namibia. Third edition*. Hawthorn, Australia : Lonely Planet. 817 pp
- Taylor, J.E., Yunez-Naude, A., Dyer, G.A., Stewart, M. and Ardila, S. 2002. *The Economics of "Eco Tourism:" A Galapagos Island Economy-wide Perspective*. University of California, Davis. http://www.reap.ucdavis.edu/working_papers/jet-galapagos.pdf [15 Apr. 2005].
- TDWG 2004. ABCD Schema – Task Group on Access to Biological Collection Data. <http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/default.htm> [Accessed 13 Apr. 2005].
- Thackway, R. and Cresswell, I. (eds). 1995. *An Interim Biogeographic Regionalisation for Australia: A Framework for Setting Priorities in the National Reserves System Cooperative Program*. (Version 4.0) Canberra: Australian Nature Conservation Agency. <http://www.ea.gov.au/parks/nrs/ibra/version4-0/index.html>. [Accessed 15 Apr. 2005].
- Thomas, C.D., Cameron, A., Gree, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., Siqueira, M.F., Grainger, A, Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Peterson, A.T., Phillips, O.L. and Williams, S.E. 2004. *Nature* 427: 145-148.
- Thompson, D.R., Furness, R.W. and Monteiro, L.R. 1998. Seabirds as biomonitors of mercury inputs to epipelagic and mesopelagic marine food chains. *Science of the Total Environment* 213: 299-305.
- Torto, B. & Hassanali, A. 1997. Progress in the search for anti-arthropod botanicals. *Recent Research Developments in Phytochemistry* 1: 475-488.
- Tsontos, V.M. and Kiefer, D.A. 2000. Development of a dynamic biogeographic information system for the Gulf of Maine. *Oceanography* 13(3): 25-30. <http://iobis.org/Plone/about/2000Tson.pdf>. [Accessed 15 Apr. 2005].
- University of Queensland. 2004. *Welcome to Lucidcentral*. Centre for Biological Information Technology, University of Queensland. <http://www.lucidcentral.org/> [Accessed 15 Apr. 2005].

- UTU-Biota. 2004. *GBIF Demonstration project 2003*. Biota BD Ltd and University of Turku. <http://gbifdemo.utu.fi/> [Accessed 15 Apr. 2005].
- Vallance, T.G., Moore, D.T. and Froves, E.W. 2001. *Nature's Investigator: The Diary of Robert Brown in Australia 1801-1805*. Canberra: ABRS.
- van Staden, V., Erasmus, B.F.N., Wingfield, M.J. and van Jaarsveld, A.S. 2004. Modelling the spatial distribution of two important South African plantation forestry pathogens. *Forest Ecology and Management* 187(1): 61-73.
- Ved, D.K. 1998. Regulating export of endangered medicinal plant species – Need for scientific rigour *Current Science* 75(4): 341-343 <http://www.ias.ac.in/currsci/aug/articles8.htm> [Accessed 15 Apr. 2005].
- Vieglais, D. 1999. The Species Analyst. Integrating Disparate Biodiversity Resources using Information Retrieval Standards (Z39.50). Powerpoint presentation. <http://www.tdwg.org/daveTDWG.htm> [Accessed 15 Apr. 2005].
- Vieglais, D. 2003a. *Species Analyst Revision 1.6*. Lawrence, KA: University of Kansas Natural History Museum and Biodiversity Research Center. <http://speciesanalyst.net/> [Accessed 15 Apr. 2005].
- Vieglais, D. 2003b. *The Darwin Core. Revision 1.5*. Lawrence, KA: University of Kansas Natural History Museum and Biodiversity Research Center. <http://speciesanalyst.net/docs/dwc/> [Accessed 15 Apr. 2005].
- Wake, D.B. 2004. *Biodiversity Informatics Approaches to Taxon-Based Studies – The Amphibia as an Exemplar*. Copenhagen: GBIF. Powerpoint presentation (10 Mb). http://circa.gbif.net/Public/irc/gbif/pr/library?!=power_point/presentations_assembly/6_wake_pps/EN_1.0_&a=d [Accessed 16 Aug. 2004].
- Wannenburgh, A. and Mabena, S. 2002. National Indigenous Forest Inventory. *National Forests and Woodlands Symposia III*. Pretoria: Department of Water Affairs and Forestry. <http://www.dwaf.gov.za/Forestry/FTIS/symp2002/inventory.doc> [Accessed 15 Apr. 2005].
- Warman, L.D., Sinclair, A.R.E., Scudder, G.G.E., Klinkenberg, B. and Pressey, R.L. 2004. Sensitivity of Systematic Reserve Selection to Decisions about Scale, Biological Data, and Targets: Case Study from Southern British Columbia. *Conservation Biology* 18(3): 655-666.
- Wassenaar, L. and Hobson, K. 1998. Natal Origins of Migratory Monarch Butterflies at Wintering Colonies in Mexico: New Isotopic Evidence. *Proceedings of the National Academy of Sciences* 95: 15436-15439.
- Watson, J.P. 1974. Termites in relation to soil formation, groundwater, and geochemical prospecting. *Soils and Fertilizers* 37: 111-114.
- Weiss, C. and Eisner, T. 1998. Partnerships for value-added through bioprospecting. *Technology in Society* 20: 481-498.
- Wells, F., Metzeling, L. and Newall, P. 2002. Macroinvertebrate Regionalisation for use in the Management of Aquatic Ecosystems in Victoria, Australia. *Environmental Monitoring and Assessment* 74(3): 271-294.
- West, J.G. and Whitbread, G.H. 2004. *Australian Botanical Informatics serving Science and Society*. Copenhagen: GBIF. Powerpoint presentation (14 Mb) http://circa.gbif.net/Public/irc/gbif/pr/library?!=power_point/presentations_assembly/5_west_pps/EN_1.0_&a=d [Accessed 16 Aug. 2004].
- Wheeler, Q.D., Raven, P.H. and Wilson, E.O. 2004. Taxonomy: Impediment or expedient? *Science* 303: 285.
- Williams, P., Gibbons, Margules, D., Rebelo, C., Humphries, A. and Pressey, R. 1996. A comparison of richness hotspots, rarity hotspots and complementary areas for conservation diversity using British birds. *Conservation Biology* 10: 155-174.
- Wiltshire, P.E.J. 2001. *Environmental Profiling and Forensic Palynology. Background and potential value to the criminal investigator* British Association for Human Identification. http://www.bahid.org/docs/NCF_Env%20Prof.html [Accessed 15 Apr. 2005].
- Winker, K. 2004. Natural History Museums in a Postbiodiversity Era. *BioScience* 54(5): 455-459.
- World Resources. 1992. [Global Biodiversity Strategy: Guidelines for action to save, study and use Earth's biotic wealth sustainably and equitably](http://www.wri.org/publications/1992/01/01_01.html). WRI, IUCN, UNEP, FAO, UNESCO. http://biodiv.wri.org/pubs_content.cfm?PubID=2550 [Accessed 13 Apr. 2005].
- Zak, D.R., Holmes, W.E., White, D.C., Peacock, A.D. and Tilman, D. 2003. Plant Diversity, Soil Microbial Communities, and Ecosystem Function: Are there any links? *Ecology* 84(8): 2042-2050. <http://www.bio.psu.edu/ecology/calendar/Zak.pdf> [Accessed 15 Apr. 2005].
- Zhou, Z.; and Pan, W. 1997. Analysis of the viability of a giant panda population. *Journal of Applied Ecology*; 34(2): 363-374.
- Zwick, P. 2003. Shapes and patterns of wingpad development in the Plecoptera, in Gaino, E. (ed.) *Research update on Ephemeroptera and Plecoptera; University of Perugia*, pp. 477- 483. <http://www.unipg.it/maystone/PDF%202001%20proc/ZWICK2%20IJM%20proceedings.pdf> [Accessed 15 Apr. 2005].

Index to Chapter 1

A

abundance, 19
Acacia, 49, 52, 76
Adelie penguin, 30
African Archaeological Database. *See Information Systems*
African-Eurasian Migratory Water Bird Agreement, 30
agricultural industry, 48
agricultural pests, 51
agricultural production, 2
agriculture, 48
Ailuropoda melanoleuca, 21
albatrosses, 26, 30, 31
alien species, 26
amphibians, 37
AmphibiaWeb. *See Information Systems*
Angle-stemmed Myrtle, 25
Anoplophora glabripennis, 16
anthrax, 58
anthropology, 75, 76
aquariums, 72
aquatic, 82
aquatic biology, 46
aquatic ecosystems, 41
aquatic invertebrates, 27
Arabian oryx, 37
archaea, 38
archaeology, 39
Argentine Ant, 27
art, 73
arthropods, 27, 51
Artificial Neural Networks. *See Software*
Asian Long-horned Beetle, 14
Asterias amurensis, 28
Australasian Bird Image Database. *See Information Systems*
Australian Bird and Bat Banding Scheme, 30
Australian magpie, 50
Australian Virtual Herbarium. *See Information Systems*
Austromyrtus gonoclada, 25
Automated identification tools, 13
avian endemism, 75
avian influenza, 57
avian malaria, 57

B

bacteria, 38, 51
badgers, 13
ballast water, 28
bamboo, 74
bark paintings, 74
Barnardius zonarius, 50
bats, 13
bees, 13
Belgian Co-ordinated Collections of Micro-organisms (BCCM), 22
BioBlitz, 70
BioCase. *See Information Systems*

BIOCLIM. *See Software*
BIODEPTH, 35
biodiversity assessment, 19, 69, 70
biodiversity surrogates, 42
biogeographic studies, 2, 14
biological control, 27, 28, 29, 48
bioprospecting, 61
bioregions, 40
biotechnology, 48, 55
bioterrorism, 57, 58
biotic surveys, 36
BIOTREE, 35
bird flu, 57
Bird observing, 79
bird strikes, 64
Bird strikes, 63
Birds-of-Paradise, 74
Bonn Convention, 29
border control, 65
botanic gardens, 72
botanical drugs, 66
Bufo marinus, 28
BumblebeeID, 10

C

Cacatua
 roseicapilla, 50
 tenuirostris, 50
Cactoblastis cactorum, 29
Caesalpinia bonduc, 74
Calabash Program, 69
Californian Condor, 58
Canna indica, 74
cash crops, 50
cassava, 48
Cassia brewsteri, 49
Census of Marine Life. *See Information Systems*
centres of endemism, 19, 20
cerrado, 32
checklists, 12
Chemotaxonomy of Xylariaceae, 61
China-Australia Migratory Bird Agreement, 29
cicadas, 13
Cicadas of South-East Asia and the West Pacific, 7
climate change, 2, 32, 52, 57
coastal dune erosion, 80
cod fish, 52, 53
codling moth, 23
complementarity, 42, 43, 44
conservation, 51
conservation assessment, 19
conservation planning, 2
conservation priorities, 42
contaminants, 54
Convention on Biological Diversity, 26, 27
Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), 65
Convention on Migratory Species, 29
Cordyline australis, 49
corella, 50
Corymbia, 10

dichromophloia, 10
umbonata, 10
Crebatulus lacteus, 22
critical habitat corridors, 43
cross breeding, 48
cultivated plants, 54
customs, 65
Cydia pomonella, 23

D

Danaus plexippus, 30
Darwin Core. *See Standards and Protocols*
data interchange, 3
DDT contamination, 58
Decision Trees, 14
dengue, 57
Depressaria pastinacella, 29
Desert Quandong, 48
DiGIR. *See Standards and Protocols*
disease vectors, 57
diseases, 57, 66
distributed data, 3, 14, 15
DNA, 62
Dreissena polymorpha, 28
drugs, 65

E

earthworms, 27
ecological communities, 19
Ecological Database of the World's Insect Pathogens.
See Information Systems
ecology, 34
ecosystem function, 35
ecosystem health, 35
ecotourism, 71, 72
ecotourism certification, 71
ecotourism guides, 71
education, 67
Eichhornia crassipes, 28
Elapid snakes, 14
Elapidae, 17
Electronic Catalogue of Names of Known Organisms
(ECat), 8
Elephants, 21
Endangered Species Program of the U.S.A., 25
endemism, 14, 19, 38
Environment Law Institute, 70
environment protection, 46
environmental contaminants, 57, 58
environmental gradients, 16
environmental modelling, 4, 14, 51, 52
environmental regionalisations, 2
Environmental Specimen Bank. *See Information Systems*
epidemiological research, 59
epidemiology, 57
equine encephalitis, 57
Eru, 49
ethnobotany, 76
Eucalyptus, 10
Eukaryotic parasites, 22
European Network for Biodiversity Information (ENBI),
22
European Pied Flycatcher, 32

evolution, 34, 37
evolutionary biology, 60
exploration, 73
Ezemvelo Nature Reserve, 45

F

Ferdinand Bauer, 73
fibres, 74
Ficedula hypoleuca, 32
field guides, 10
Fire Ant, 25
fisheries, 48
fishery production, 2
fishing
 recreational, 78
fishing, 65
fishing bycatch, 54
floras and faunas, 8
food industry, 48
food processing, 48
forensic entomology, 63
Forensic entomology, 63
forensic science, 62
forest production, 2, 51
forestry, 48, 51
Frenchie beetle, 28
freshwater fisheries, 53
FrogLog, 26
Frogwatch, 70
fructose production, 49
fungi, 51
fungus species, 51

G

galah, 50
GAM. *See Software*
GAP Analysis Program, 36
Gardening for Biodiversity, 79
GARP. *See Software*
GBIF. *See Organizations*
GBIF Demonstration Project, 4, 10, 15, 36
GBIF Portal. *See Information Systems*
GenBank, 45, *See Information System*
Gene fragments, 62
gene transfer, 48
Generalised Additive Models (GAM), 16
Generalised Linear Models (GLM), 2, 14, 16
Genetic Algorithm for Rule-set Production. *See*
Software: GARP
genetic breeding, 48
genetic improvement, 49
genetically modified crops, 58
genetically modified organisms (GMOs), 58
genomes, 38
genomics, 37
Geographic Information Systems, 15
Giant Cane Toad, 28
Giant Panda, 21, 43
Giant River Prawn, 53
Giant-petrel, 26
Global Register of Migratory Species. *See Information*
Systems
GLOBE, 67

Gnetum
africanum, 49
buchholzianum, 49
goats, 25
Google Images, 69
grasshoppers, 13
Gray Card Index, 8
Grey backed cane beetle, 28
Guanacaste Conservation Area, 13, 45, 68
guidebooks, 71
Gulf of Maine Biogeographic Information System. *See*
Information Systems
Gymnobilideus leadbeateri, 21
Gymnogyps californianus, 58
Gymnorhina dorsalis, 50

H

habitat fragmentation, 35
habitat loss, 35
hanta viruses, 59
harvesting of wild populations, 49, 50
health, 57
 environmental, 57
 human, 57
heavy metal concentrations, 59
Helicobacter pylori, 38
herbal medicines, 60
HIV, 57
Holocene climates, 37
Homalodisca coagulata, 27
honey, 50
host specificity, 22
hunting, 78
HymAToL, 7

I

ICLARM, 7
illegal bird trade, 65
image databases, 12, 69
Imagens da Biodiversidade Brasileira, 69
Index Fungorum, 8
Index of Viruses, 8
indices of diversity, 14
indigenous art, 74
infectious diseases, 58
Information System
 GenBank Database, 38
Information Systems
 African Archaeological Database, 39
 AmphibiaWeb, 18, 26
 Australasian Bird Image Database, 69
 Australian Natural Resources Atlas v. 2.0, 18, 34
 Australian Plant Pest Database, 66
 Australian Virtual Herbarium, 3, 9
 BioCase, 3
 Bird Remains Identification System (BRIS), 64
 Bird Strike Information System (IBIS), 64
 Census of Marine Life, 82
 Digital Orthoptera Specimen Access (DORSA), 69
 Ecological Database of the World's Insect Pathogens,
 51
 Environmental Specimen Bank, 58, 59
 GBIF Portal, 3, 8, 32, 59, 76

Global Register of Migratory Species (GROMS), 30
Gulf of Maine Biogeographic Information System
(GMBIS), 53
Gulf of Maine Biogeographic Information System
Atlas (GMBIS), 82
Insect Identification and Biosystematic Service, 51
Intelligent Bioacoustic Identification System (IBIS),
13
MaNIS, 3
North Australian Frogs Database System, 68
Ocean Biogeographic Information System (OBIS), 82
speciesLink, 3
Tree of Life, 70
insect herbivores, 22
insects, 51
Integrated Catchment Management, 70
Integrated Taxonomic Information System (ITIS), 8
Intelligent Bioacoustic Identification System. *See*
Information Systems
Intelligent Bioacoustic Identification System (IBIS), 13
Inter-American Environment Program, 70
International Plant Name Index (IPNI), 8
invasive species, 26
inventories, 10
irreplaceability, 42
IUCN Red List of Threatened Species, 25

K

kangaroos, 50

L

land resources, 46
landscape restoration, 48
landuse planning, 46
Lantana, 29
Lassa fever, 59
lead contamination, 58
Leadbeater's Possum, 21
Leptospermum, 18
Leucaena, 52
Linepithema humile, 27
Linnaeus II. *See Software*
Lucid. *See Software, See Software*
Lyme disease, 59

M

Macrobrachium rosenbergii, 54
macroinvertebrates, 41, 46
macropods, 50
Macropus fuliginosus, 50
malaria, 57
Man and the Biosphere Programme, 45
Manihot esculenta, 48
MaNIS. *See Information Systems*
marine, 41, 82
mealybug, 28
megafauna, 37
mercury contamination, 58
microbial diversity, 38
migration, 57
migratory species, 29

Millenium Seed Bank, 45
mining, 55
mites, 51
mollicutes, 51
Monarch Butterfly, 30
Museum School Partnership program, 67
museums, 72
Museums around the world, 72
mycorrhiza, 55
Mycteria americana, 23

N

National Zoo Biodiversity Monitoring Project, 67
natural resource management, 46
nematodes, 51, 66
nemertean, 22
Neotropical species distributions, 4
New Endeavour, 69, 73
Nickernuts, 74
nomenclature, 75
North American Hunting Heritage Accord, 78
North American Wood Stork, 23
Northern pacific seastar, 28
Nothofagus cunninghamii, 14, 37

O

Ocean Biogeographic Information System. *See*
Information Systems
on-line identification tools, 69
Opuntia, 29
orchid cultivation, 55
orchids, 65
Organizations
 Albufera International Biodiversity Group (TAIB), 47
 American Board of Forensic Entomology, 63
 American Museum of Natural History, 37
 Arizona State Museum, 39
 Australian Biological Resources Study (ABRS), 11,
 17
 Australian Broadcasting Commission, 39
 Australian Department of Environment and Heritage,
 25
 Australian Museum, 17, 69
 Australian National Botanic Gardens, 55
 Australian National Land and Water Resources, 43
 Binatang Research Centre, Papua New Guinea, 13
 Biodiversity Research Center, 58
 Birdlife International, 20, 47
 Bishop Museum, 68
 Bureau of Flora and Fauna, 17
 Centers for Disease Control and Prevention (CDC),
 58
 Centre for Plant Biodiversity Research (CPBR), 9, 11
 Centre for Resource and Environmental Studies, 21
 Chaing Mai University, 75
 Chinese Academy of Sciences, 45
 Cleland Conservation Park, 45
 Comisión Nacional para el Conocimiento y Uso de la
 Biodiversidad (Conabio), 29, 58
 Conservation International, 20
 Convention on Biological Diversity (CBD), 76
 Council for Biotechnology Information, 61
 Duke University, 67

Durrell Wildlife Conservation Trust, 72
Electric Power Research Institute, 80
Environment Canada, 25
Environmental Resources Information Network
(ERIN), 3
European Centre for Nature Conservation, 75
European Molecular Biology Laboratory (EMBL), 38
European Union for Bird Ringing, 30
FAPESP-Biota, 61
Federal Geographic Data Committee, 34
Field Museum, 69
Food and Agriculture Organization of the United
 Nations (FAO), 53
Global Biodiversity Information Facility (GBIF), 4,
 84
Global Invasive Species Program (GISP), 27
Institute for Comparative Genomics, 37
Institute of Amazonian Research, 4
International Development Research Centre, 51
International Union for the Conservation of Nature
 and Natural Resources (IUCN), 25
Inwent Zschortau (Leipzig), 71
Jersey Zoo, 72
Jurong Bird Park, 72
Kirstenbosch National Botanical Garden, 72
Laboratory of Ethnobotany, 76
Lankester Botanical Gardens, 55
Maryland Department of Natural Resources, 46
McClung Museum, 74
Monterey Bay Aquarium, 72
National Biodiversity Institute (Inbio), 61, 68
National Biodiversity Network (NBN), 68, 70
National Centre for Integrated Pest Management, 51
Natural History Museum, 67
New York Botanic Gardens, 69
New Zealand Department of Conservation, 25
North Carolina Nature Museums and Science
 Centers, 69
Queensland Museum, 59
Queensland Parks and Wildlife Service, 25
Royal Botanic Gardens Kew, 44
Royal Horticultural Society, 79
San Diego Zoo's Wild Animal Park, 45
Saskatchewan Agriculture, Food and Rural
 Revitalization, 50
Society for Growing Australian Plants, 54, 79
South African Institute for Natural Resources, 46
South Lakes Wild Animal Park, 45
Swedish Museum of Natural History, 58, 59
Taxonomic Databases Working Group (TDWG), 3
The Natural History Museum, 70
The Nature Conservancy, 71
The World Bank, 46
The World Conservation Union (IUCN), 44, 46
U.S. Fish and Wildlife Service, 25
Unit for Social and Environmental Research, 75
United Nations Environment Programme (UNEP), 71
United States National Parasite Collection (USNPC),
 59
University of Kansas, 58
University of Toronto, 22
University of Turku, 4
University of Waterloo, 47
US Environment Protection Authority (EPA), 46, 47
Washington University in St. Louis, 39
WWF Guianas, 66
Xishuangbanna Tropical Botanical Garden, 67

Zooarchaeology Laboratory Comparative Vertebrate Collection, 39
 Ornamental Plants database, 54
Oryza
glumaepatula, 48
longistaminata, 48
nivara, 48
rufipogon, 48
 Overfishing, 53

P

palynology, 63
 Parasite Database, 22
 parasites, 59, 60
 parasitology, 59
 parataxonomists, 13, 68, 69
 parsnip, 29
 parsnip web worm, 29
Pastinaca sativa, 29
 pathogenic bacteria, 38
 pathogens, 51, 57
 Pattern Analysis, 19
 pattern recognition, 13
 pests, 66
 petrels, 31
 pets, 66
 pharmaceuticals, 61
 phenology, 23
 Phylogeny, 7
Phyophthora cinnamomi, 25
 phyto-mining, 55
 phytoremediation, 55
 Plant Genome Databases, 37
 plantation forestry, 51, 52
 Platform Terminal Transmitters, 31
 pollen, 63
 pollution monitoring, 55
 Population Viability Analysis, 21
 populations, 19
 Port Lincoln Parrot, 50
Prasophyllum petilum, 37
 protected areas, 43
 protozoa, 51
 provenances of cultivated species, 49, 51, 52
 Przewalski horse, 44
 public outreach, 67
 public participation conservation programs, 69
 public safety, 57
 Publications
 A Guide to the Birds of Panama, 72
 A Mexican case-study on a centralised data base from World Natural History Collections, 76
 Acacia in Australia: Ethnobotany and Potential Food Crop, 76
 Acacias of Australia, 12
 Amazonian Biodiversity Estimation, 42
 Arthropods of Economic Importance, 11
 Atlas Florae Europaeae, 15
 Atlas of Australian Birds, 15
 Atlas of Elapid Snakes of Australia, 15, 16, 17
 Atlas of the Birds of Mexico, 15
 Atlas of the British Flora, 14
 Atlas of Vertebrates Endemic to Australia's Wet Tropics, 16
 AusGrass, 12

Australian Mammal Audit, 12
 Australian Plant Collectors and Illustrators 1780s-1980s, 73
 Australian Plant Image Database, 69
 Australian Plants online, 79
 Australian Terrestrial Biodiversity Assessment, 18
 Australian Tropical Rainforest Trees and Shrubs, 12
 Bats of the Indian Subcontinent, 11
 Biodiversity Toolbox for Local Government, 42
 Biodiversity World, 42
 Bioinformatics: Sequence, Structure and Databanks – A Practical Approach, 38
 Birds of Argentina and Uruguay, 10
 Birds of Europe, 11
 Butterflies of Australia, 10
 Butterflies of North America, 10
 Canada's National Marine Conservation Areas System Plan, 41
 Catalogue of the Chalcicoidea of the World, 11
 Catalogue of the species of the Annelid Polychaetes of the Brazilian Coast, 10
 Census of Australian Vascular Plants, 15
 Checklist and distribution of the liverworts and hornworts of sub-Saharan Africa, 12
 Checklist of Amphibian Species and Identification Guide for North America, 12
 Checklist of Online Vegetation and Plant Distribution Maps, 34
 Checklist of the Amphibians and Reptiles of Rara Avis, Costa Rica, 12
 Checklist of the Ants of Michigan, 12
 CITES Identification tools and Guides, 65
 Costa Rica's National Parks and Preserves, 72
 Crabs of Japan, 11
 Davalliaceae, 11
 Diptera species pages, 70
 Distributions of Mexican birds, 17
 Dragonfly Recording Network, 10
 Ecotourism Training Manual for Protected Area Managers, 71
 Endemic Bird Areas, 20
 Environmental Contaminants of Amphibians in Canada, 59
 Ethnobotany: Plants and People Interacting, 76
 Eucalypts of Southern Australia, 12
 Evolution and Mass Extinction, 37
 Farming Freshwater Prawns, 53
 Fauna Malesiana, 11
 Fauna of New Zealand, 8
 FaunaItalia, 8
 Fife Bird Atlas, 14
 Fishes of the North-Eastern Atlantic and Mediterranean, 11
 Flora of Australia online, 8
 Global 200 Ecoregions, 41
 Handbook for Botanic Gardens on Reintroductions of Plants to the Wild, 44
 History of systematic botany in Australasia, 73
 Indian Medicinal Plants, 66
 Interim Biogeographic Regionalisation of Australia (IBRA), 40
 Interim Marine and Coastal Regionalisation for Australia (IMCRA), 41
 John Gould's Birds of Asia, 73
 Key to Common Chilocorus species of India, 11
 Key to Cotton Insects, 11
 Lewis and Clarke Expedition, 73

Long-term Monitoring of Australia's Biological Resources, 47
 Millennium Atlas of Butterflies in Britain and Ireland, 14
 Mites in Soil, 12
 Modelling Forest Systems, 52
 Mosquito-borne diseases, 57
 Moths of North America, 15
 National Forestry Programme for Swaziland, 51
 National Vegetation Information System (NVIS), 34
 Natural resource management and vegetation – an overview, 46
 Nature's Investigator: The Diary of Robert Brown in Australia 1801-1805, 73
 New Biogeographic Regionalisation for Tasmania, 41
 New Wildlife ID Manual, 66
 Ontario Herpetofaunal Summary Atlas, 14
 Papua New Guinea Conservation Needs Assessment, 43
 Papua New Guinea Country Study on Biological Diversity, 42
 Phanerogamic Flora of the State of São Paulo, 8
 Plant-derived Drugs: Products, Technology, Applications, 61
 Protea Atlas, 15
 Rabbit Haemorrhagic Disease: Issues in Assessment for Biological Control, 57
 Reference List for Plant Re-Introductions, Recovery Plans and Restoration Programmes, 44
 Regional Land Use Plans and Land Resource Management Plans (LRMPs) in British Columbia, 46
 Rights-of-Way Environmental Issues in Siting, Development and Management, 80
 Species Decline: Contaminants as a Contributing Factor, 18
 Species Identification and Data Programme, 53
 Species Richness bibliography, 19
 Spiders of Australia, 12
 Stag Beetle Biodiversity Action Plan, 67
 The Age of the Megafauna, 39
 The Green Legacy, 44
 The New Atlas of Australian Birds, 15
 The Weight of a Petal: The Value of Botanic Gardens, 44
 The World of Insects in Chinese Art: A Special Exhibition of Plant-and-Insect Paintings, 74
 Threatened Species Recovery Plans, 25
 Tools for Assessing Biodiversity Priority Areas, 42
 Training Manual for Community-based Tourism, 71
 Tree of Life, 12
 Tuna Bycatch Action Plan, 54
 UK Habitat Classifications, 34
 Valuing Ecotourism as an Ecosystem Service, 71
Pultenaea, 9

Q

quarantine, 66

R

rabbit calicivirus, 57
 rabbit haemorrhagic disease, 57
 rainforest inventories, 10

rainforest trees, 22
 rapid biodiversity assessment, 42
 rapid ecological assessment, 42
 regional planning, 41
 relative abundance, 14
 replication, 42
 representativeness, 42
 reptile diversity, 18
 reserve selection, 2, 42
 reserve-selection algorithms, 44
 ribosomal RNA sequence analysis, 38
 rice, 48
 risk assessment, 58
 roads and services, 80
 root-rot fungus, 25

S

Saltcedar, 27
Santalaum acuminatum, 48
 satellite tracking devices, 30
 school level education, 67
 screening for bioactive compounds, 61
 sea turtles, 30
 seahorses, 66
 sericulture, 50
 Shahtoosh, 65
 shells, 74
 Smithsonian National Museum of Natural History, 72
 snake antivenom, 59
 snakebite, 59
 social uses of biodiversity, 75
 Software
 Artificial Neural Networks, 14
 Australian Heritage Assessment Tool, 19
 Automatic Bee Identification Software (ABIS), 13
 BIOCLIM, 2, 14, 16, 17, 37
 BioRap, 42
 DELTA, 11
 DOMAIN, 16
 empirical forest models, 52
 EstimateS, 19
 GAM, 14
 GARP, 2, 14, 16, 17
 IntKey, 11
 Lifemapper, 68
 Linnaeus II, 11
 Lucid, 11
 PATN, 19
 PoliKey, 11
 process-based forest models, 52
 RAMAS, 21
 Species Analyst, 3
 VegClass, 34
 VISTR, 36
 WorldMap, 19
 XID Authoring System, 11
Solenopsis invicta, 25, 26
 Southern Elephant Seal, 21
 spatial patterns, 21
 Species Analyst. *See Software*
 species declines, 18
 species density, 19
 species distribution atlases, 14
 species distribution modelling, 3, 14, 16, 57
 species distributions, 14

species diversity, 19
species extinctions, 37
species modelling, 18
species richness, 19
species translocation, 2, 26
Species2000, 8
speciesLink. *See Information Systems*
specimen loans, 4
Spondias mombin, 49
Standards and Protocols
 Darwin Core, 3
 DiGIR Protocol, 3
 HISPID, 3
 Vegetation Classification Standards, 34
 Z39.50, 3
street trees, 80
survey planning, 36
sustainable forestry management, 51
Sydney Parkinson, 73

T

Tamarix ramossissima, 27
Tarengo Leek Orchid, 37
Tasmanian Shy Albatross, 31
taxonomic research, 7
Taxonomic Search Engine (TSE), 8
taxonomy, 7
TDWG. *See Organizations*
Terminalia alata, 55
termites, 50
terrorism, 57, 58
Threat Abatement Plans, 25
Threatened Species Program, 25
Threatened Species Recovery Plans, 25
threats to endangered species, 25
Toad Action Group, 67
traditional use, 48
Tree of Life, 7
tree roots, 80

Tropicos, 8
Tropidechis carinatis, 17
turtles, 54, 65

U

USGS-NPS Vegetation Mapping Program, 34

V

vegetation, 34
vegetation mapping, 34
virtual reference systems, 69
viruses, 51

W

Walking with Woodlice, 70
water quality, 46
water resources, 46
Waterwatch, 67
weed invasions, 27
West Nile virus, 57
western grey kangaroo, 50
wild relatives of cultivated crops, 48
wildlife parks, 45, 72
wildlife trade, 65, 66
woodland birds, 35
wool, 74
World Federation of Culture Collections (WFCC), 22

Z

Z39.50. *See Standards and Protocols*
Zebra Mussel, 28
Zoological Museum of Amsterdam, 7
zoos, 72

Chapter 2

Initiating a Collection Digitisation Project

Section 1 Introduction	1
Purposes of this document.....	1
What do we mean by digitisation?.....	2
Target audience & scope of this document.....	2
Document Overview.....	3
Section 2 Why should we digitise our collections?	3
Section 3 What you should do before you get started.....	5
Planning is important.....	5
Identify your goals.....	6
What are your current limitations and resources?.....	11
Produce a Tentative Business Case	17
Pick a database solution.....	18
Develop an Action Plan.....	18
Running the project.	24
Section 4 Organising Information and Representing Data	26
Object Information (as you see it)	26
Reference and Ancillary Information	28
Data (as the computer sees it).....	28
Data Views and Storage Formats	32
Standards	33
Data Quality.....	35
Language	39
Intellectual Property Rights.....	39
Section 5 The Data Model.....	39
Introduction	39
The Base Unit of a Simple Data Model.....	40
The Focus of a Data Model	41
Modular Design of the Data Model: The Information Management System	45
Implementing a Data Model.....	46
Data model solutions vs. programmatic tools.....	48
Complexity	48
Section 6 Deciding on a particular database solution	51
Which database(s) should you use?.....	53
What are characteristics of a good database solution?.....	55
Does it have the right focus and features for you?.....	60
References	65
Appendix A: Business Case Considerations.....	69
Why digitise your collection?.....	69
Identify your goals.....	69
Appendix B: Action Plan Issues	69

This Chapter is equivalent to:

Frazier, C.K., Wall, J., and S. Grant. 2008. *Initiating a Natural History Collection Digitisation Project*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 75 pp.
ISBN: 87-92020-05-4 (available as a standalone PDF from <http://www.gbif.org>)

Section 1: Introduction

Purposes of this chapter

Deciding to undertake a digitisation project can be a daunting process. By now, you've probably already noticed that "everybody is doing it." You've probably developed a strong suspicion that you are in danger of getting left behind in this information age if you don't start something soon. Maybe you're being pressured from one or more user groups, your administration, or colleagues to provide electronic data. Or maybe your main reason for showing an interest is that you've run up against real shortcomings in fulfilling your mission that you feel could be addressed better if your collection were digitised.

Regardless of your collection size or your reasons for wanting to start a digitisation project, your core interest probably isn't learning a huge amount about bioinformatics, system design, or information theory. You want to get it done, but where do you start? What do you really need to know in order to find a solution that meets your needs and resources without having to get another degree in computer sciences along the way?

There is simply no getting around the fact that you will have to consider a number of factors and make a large number of decisions if you want to get a workable solution that meets your needs. We know from experience that the first step most people go through when they are starting out is to think they'll just do what everybody else is doing. If only it were that simple. There are currently dozens of solutions in use, some home-grown, some commercial, some open-source. Some are tailored to your kind of situation, some are more generalized but customizable, and virtually all are going to be limited in some way that is ultimately important to you. Some are complex and powerful, some are simple, some are well-documented and some are not. And it's not just enough to know what is out there because some people or institutions are really happy with what they use, some are not, and most are somewhere in the middle.

This document is designed to give you the confidence to get started and to help you make the right decisions as you plan a digitisation project. We, the authors, have years of experience working with collections large and small and we have tried to instil this into this document so you can more efficiently ask the right questions and make the appropriate plans prior to committing your resources. Our goal in this paper is to provide you with the information that we wished we had at the outset of our digitisation projects. We've tried to give you the information that you might not get by talking to representatives or advocates of a particular solution.

Both information technology and informatics theory are evolving rapidly and finding the right solution for your needs is something of a moving target. Specific solutions that were in vogue just a few years ago now seem quaint or antiquated. In some cases, there are new possibilities on the horizon that might seem just what you need, but will they deliver as promised? Should you invest your time perusing the latest conference proceedings rather than reading through this document?

We have strived to make this document more relevant and less time-dependent by focusing on a simple fact we have learned about all IT systems. No matter how powerful, cutting-edge, or expensive they might be, an IT system is only as valuable as the quality of the data it contains. In the future or now, you will want to find a solution that allows you to enter, maintain, and output quality information. Thus, our emphasis in this document is to help you

determine how to do that given the resources you have and your particular goals. This, in itself, is an extensive topic with many ramifications which may not be anticipated by those just starting out. There are several areas which have already been well covered in previous works and, rather than repeat those discussions, where appropriate we provide suitable references that should be consulted by those needing fuller understanding of the topic. The ideas exposed in this paper are the best we can come up with given our current experience working with a variety of digitisation projects and we hope they are helpful to those researchers in a similar position to ourselves.

Though in some cases we will recommend particular solutions in this document, overall our intent is not to provide a set of “best practices.” It is our opinion that the appropriate solution for a specific institution or collection is highly dependant on the unique circumstances of each institution, so no single best practise exists. Getting the correct answer for yourself is the first step in implementing your own successful project.

What do we mean by digitisation?

Digitisation refers to the capture of information in electronic form. The basic information of concern to our community comes from checklists, field notebooks, collected specimens or may be extracted from publications, documents or other media. The basic unit of information may concern physical objects, like specimens in an herbarium, or events, such as observations of birds singing in a forest on a particular day or the collection of a number of organisms in a pitfall trap left overnight. The results of digitisation can be stored or expressed in a variety of ways such as formatted or marked-up documents, spreadsheets, web pages or web sites, flat-file or relational databases, maps or GIS systems.

Another important distinction for our community is that digitisation may refer to the electronic capture of an *image* of an object or it can also be used to refer to the capture of *textual information* about an object or extracted from an object that contains text. While both of these may be referred to as digitisation, we prefer the term “imaging” for the former. The specifics of imaging specimens is not dealt with in depth in this document. A good source of information on imaging is the Global Biodiversity Information Facility (GBIF) paper “Digital Imaging of Biological Type Specimens: A Manual of Best Practice” (Häuser et al., 2005). The term “databasing” is preferred to describe the process of capturing text-based information. In addition, many digitisation projects involve creation of an information system that holds both images and text-based information. In this case, “databasing” or “digitisation” refers to the creation of this system to hold information.

Target audience & scope of this chapter

This document is aimed at anyone who is interested in starting a digitisation project for the first time. It is hoped that it will also be useful to more experienced digitisers as well. The central focus of this document is to assist managers of specimen-based biological collections. Even so, it is recognized that managers may want to digitise their collections for a variety of reasons and with different goals for their digital information. Some, for example, may simply want to create an electronic catalogue of what they have, but others may be trying to create detailed records of their collection events, and still others may be looking for a way to integrate their voucher collections into an information system supporting project-based research at their institution. In this document, we provide information that we hope will be useful in planning a digitisation effort regardless of the specifics of one’s particular goals.

Digitisation projects can vary in size from a single person digitising a single specimen to an institution wide program recording millions of specimens. Many of the issues discussed in this document are independent of the size of the collection or scope of the digitising effort. In other cases, we seek to provide guidance specific to the scope of the effort.

This document does not assume any particularly high level of literacy in database design, computer technology or natural history. However, in some cases, the reader may be directed to ancillary documents to help promote a better understanding of the concepts discussed here.

Chapter overview

The next Section “Why should we digitise our collections” is a short introduction to the reasons for undergoing a digitisation effort and the benefits you can expect to obtain. Section 3, “Before you begin” is broken down into several major parts. The first four sections are concerned with setting out your goals and analysing your current situation. Once you have done this, you can then begin to consider the specific details of how you are going to implement your project, covered in ‘selecting an appropriate database’ and ‘writing a tentative action plan’. The final part addresses putting your plan into action. Section 4 discusses information (as you see it) versus data (as the computer sees it). This distinction is used to explain how the information you want the system to hold will have to be manipulated in order to be entered and maintained in a database. This Section also discusses other data issues including standards, data quality, language and intellectual property rights. Section 5 discusses the concept of the data model, starting with simple catalogues and ending with the modular design of information management systems. The Section then continues into how data models are implemented in computer systems and discusses some of the basic issues involved with implementation. Section 6 concludes with more detailed discussion and advice toward how to evaluate and select a particular database solution that will meet your needs. After having read through these Sections, the reader is encouraged to begin building the business and action plans to coordinate the development process. Appendices A and B include short outlines of the processes for developing these plans.

Section 2: Why should we digitise our collections?

Before computers, natural history collections were the physical databases from which information could be laboriously transcribed (Lane, 1996). Unless the data were then published, transcription made the data available to only one person, the researcher involved. With the advent of computers, and increasing access to data via the Internet, new ways of utilising the potential of your natural history collections has become possible.

Digitisation can have a lot of benefits for your collection, for your staff and workflow and for the potential users of your data. Nevertheless, digitisation incurs a real cost and it is important to understand and address, explicitly, the reasons why you are undergoing a digitisation effort. It may not be obvious to administrators or potential funding sources why the effort and expense are justified unless you give specific reasons. After the project is underway or completed, these reasons can then be used to generate benchmarks and other criteria used to evaluate the efficacy of the digitisation project.

An extensive review of the uses of digitised data can be found in Chapter 1 of this *Manual*, which is based on Chapman (2005c). Some common reasons for embarking on a digitisation project include:

Wider dissemination of data

Primary specimen information is typically restricted to the data embodied on the specimen sheet itself. It can therefore only be made available to whoever currently holds the specimen. Without digitisation, passing data between institutions requires either a personal visit from the interested researcher or the specimen must be loaned out, at a potentially high cost in transport and curatorial activities. Digitised data can be disseminated in many ways, primarily using the Internet, enabling many more people to access and utilise the data.

Enable your data to be studied in different ways

Once you have digitised your collection, you can then query the data in ways that were not easy to do before. For example, you can arrange the data by collector and collection date, allowing you to track the progress of collecting trips. Trying to do this in a collection arranged by family is virtually impossible. Digitised specimen records also play an important part in the estimation of species diversity (Meier and Dikow, 2004; Chapman, 2005c). So long as the relevant data is recorded in a well structured database you have the potential to view it in whatever manner you require.

Enhance curatorial activities

Digitising your collection can aid the day to day activities in your institution, usually by reducing the amount of book keeping required. It can also keep track of the status of the collection by tracking the loan status of a specimen. The quality of the collections is enhanced by identifying inaccuracies; 'lost' specimens may be rediscovered (Peterson, 2002) and standardising the terminology used on the specimen labels. Digitising quickly illustrates the absence of useful data, typically when you are trying to study the data in different ways as mentioned in the previous point. Few other activities will enable you to get to know the true depth of your collection as digitisation will (Peterson, 2002).

Protect your specimens

Digitisation inevitably requires some referencing to the original specimen. Once this is done, there is a reduced requirement to handle specimens, as the specimen data can be transported instead. By reducing handling you will increase the longevity of the original article. This is especially important for irreplaceable items such as type specimens. This does not, however, preclude nor should cause a restriction of access to the specimen, as many forms of research will still require physical examination of the item itself. Digitisation, even if you include an image of the specimen, can only reduce the frequency specimen handling, it cannot replace it. The digital collection also acts as a form of disaster management. Should the worst happen and the original collection be destroyed, the digital collection will continue to provide a valuable resource.

Aid research by reducing future transcription time

Once the specimen data is transcribed it need not be repeated for future projects which also feature the same specimen. This allows later projects to be more efficient, reducing the cost requirements.

Raising institution/collection profile

Institutions are interested in being able to access data from a broader range of sources than merely their own collections. There is also increased pressure to allow greater access to the institutions collections, resulting in improved resources (financial or otherwise) for research projects. Many new projects require online access to the resulting data, pushing forward the

creation of digital data. User appreciation of good quality data leads in turn to better appreciation of the original collection (Lane, 1996), which enhances the importance of your collection. Digitising also enables you to monitor the size, growth and usage of your collection, which is very useful when pursuing funding for new projects. Digitisation can also satisfy the CBD requirement to repatriate specimen data to the originating country (Meier and Dikow, 2004).

Enhances the ability of the institution to contribute in areas beyond its traditional remit

Traditionally, natural history institutes have acted to preserve specimens and aid researchers in nomenclatural research. When specimen data is made available, it need not only be used to supply taxonomic researchers, new areas of interest can be catered for. Data could be used in education or to increase the general public understanding of the work done by the institution. The data can also be analysed to identify gaps in the collection and create collection guides to aid future collecting trips. Many more potential uses for the data exist, which can be easily implemented once the digitised data is available.

Legislation

Making data widely available is increasingly required in some countries 'access to information for publicly funded institutions' legislation.

Section 3: What You Should Do Before You Get Started

Planning is important

Clear planning is vital to deliver a suitable database. In practical terms, setting up a digitisation effort at any scale is a project in and of itself. Application of formal project management techniques will improve the probability that a given project is successful. It is highly recommended that the implementation of a digitisation program follows the principles of good management; however a thorough discussion of project management techniques lies beyond the scope of this paper. Project management texts are widely available and it is recommended that the reader consult those before initiating a project. In this paper, we will discuss three topics that directly derive from project management. These are the business case, the action plan and risk analysis.

The business case sets out what you wish to do and establishes the benefits you expect to gain from undertaking the suggested work. It also includes an assessment of the resources required to implement the project as well as identifying which resources are currently available. Any shortfall in resources should be clearly identified and the associated costs be stated. Setting these facts out in a single document allows a clear judgement to be made on the feasibility of the project and helps to clarify why limited resources (even if only one person) should be used on digitisation.

The action plan details how the business case should actually be implemented. It contains practical information such as how many computer(s) and database(s) will be required. It also considers the number of staff; training and how the digitisation work will proceed (commonly referred to as workflow). The action plan also details where funding should be sort to cover the resource shortfall.

Risk analysis documentation is a part of the action plan which aims to consider what to do if something goes wrong. Simple examples include what will happen if a computer fails or if funding is not secured for part of the project. Consideration also goes into how to minimise the risk of an event happening. Regularly backing up the data, seeking alternative funding, having a spare computer available are all simple ways of risk mitigation considered in the risk analysis documents.

It is possible that the business case outlines an overall goal that is too large to be completed in any one project and so is broken down into several smaller projects with their own business cases, action plans and risk analyses. This is perfectly acceptable practise and the action plan for the overall business case should then outline the separate projects and how they link together to provide the overall goal. Working in this way allows large and often long term goals to be achieved in small stages without losing sight of the overall vision.

It is important not to rush the planning phase as correcting problems once the database has been released can be both difficult and time consuming. For an institution, the planning phase could easily take six months to a year to implement correctly. This may seem discouraging at first, but taking the time to properly understand your requirements will avoid disappointment when an inappropriate package is rolled out.

It may be the case that a short term solution may need to be in place before the planned database is ready. In this case your plans should include time to migrate the data from the short term database to your permanent solution; otherwise the short term solution may become your de facto permanent database!

Identify your goals

The general principles of why a digitisation project should be undertaken have been identified. While these are good general reasons, it is important to identify exactly why your specific digitisation project should be undertaken. This section aims to raise appropriate questions that every digitisation project needs to answer at some point during its lifetime. Many projects have not considered all of these questions before the project actually began, typically causing substantial additional work when changes to the project have been required. Once the questions in this section have been answered, you should have a much better idea of your projects resources, requirements and restrictions.

Institutional versus individual

Does your project cover the entire institution or is it just a one-man project? Acknowledging the scale of your project defines many of the restrictions you will face when the project goes live. For example, if your project only involves one person, the required workflow, computing and physical space requirements commitments are much smaller than for a full institutional system. Similarly a small project can only digitise a relatively small number of specimens compared to a whole institution. For an institutional project, it is very important to get the staff on board, so careful attention must be applied to the way the project interacts with the day to day work in the institution. Training and education on the project's aims and procedures are vital; if these processes are not followed from the start the chances of your project being successfully completed are significantly reduced. As the scale of the system increases the underlying database tends to have to become more complicated and more strictly designed, with less room for ad-hoc creation of new fields to respond to the needs of individual researchers.

Who are the principal clients of your solution?

There are many possible users of your data. In terms of the initial data generated from the specimen, users can include taxonomists, managers, researchers, technicians, collectors, environmentalists, non-governmental organisations, pharmacologists and the general public (Chapman, 2005a; see also Chapter 3 of this *Manual*). Inevitably there will be a small group of target users who are your principal audience. Typically for a natural history digitisation project the target audience will be one of the following:

- Individuals on a specific project,
- Researchers generally, and
- Curatorial staff at the institution.

Single person systems are usually very simple to implement and can easily be configured to the unique requirements of that individual. They usually only require a simple form of data entry and a querying system designed to produce a single target result such as a paper, flora or checklist. However, such datasets may be of limited use beyond the initial project unless a specific effort is made to make the dataset widely available. At an institutional level, it is often useful to impose a requirement on internal projects that all recorded data is made available to a central location. This thereby allows others to find previously recorded data and reduces duplication effort across projects.

Researchers may either be external or internal to the institution. Each of these two broad categories of researchers will require some form of interface to be able to access the data. External researchers will typically use a web site to access the information. Internal researchers may also use the same web site the external users have, but you may wish to allow them to access additional information (the location of the specimens in the cupboards would be a simple example). Concentrating on making data available to all rather than individual researchers has two useful consequences at an institutional level. Firstly, the data will need to be standardised across research projects this also helps guarantee long term storage of the data, making the data available to future researchers. Funding bodies often require projects to have some form of information dissemination objective. Planning to make the data available to external researchers will generally serve to fulfil this requirement. One way of making data widely available is to join a data provider such as the Global Biodiversity Information Facility (GBIF). GBIF has its own specialised form of connecting to the database which would need to be implemented.

Digitisation can also be used to assist best practice in the institution, particularly in the areas of loans and accessioning. In order to be able to do this, the staff will need a system that is available within the institution but there is no requirement for a globally available interface.

Of course, it is entirely possible that you will wish to enable many other users to access the database, each with their own special requirements which may include additional information to be captured by your digitisation effort.

Do be aware that the more target audiences you wish the database to serve, the more complicated the database will become. In the case of the clients above, each type of client needs to be able to access the data in a slightly different way, possibly requiring 3 different interfaces (data entry, internal access, external access) to be built. Take the time to discuss specific requirements with several representatives of each client group. This will help ensure that your goals match up with the needs of your target audience.

What language(s) will you support?

The more languages you need to present information in, the more complicated your dataset and interface will inevitably become by a significant margin. For example, the interface alone will have to be presented in more than one language and the data will have to be translated into each separate language. At the very least, your database system should be able to handle unusual (in English) characters such as diacritical marks.

How much data?

While this can be a manageable task for small collections it becomes progressively more difficult as the size of the collection grows and eventually digitising all specimens can become impracticable except as a very long term goal. Targeting specific parts of a collection is frequently a better strategy. This can vary based on the immediate requirements of the institution but typically will focus on easily defined groups such as families or a specific geographic area. One valid technique for digitisation is to take the most important specimens (usually the types) and concentrate the digitisation effort there.

Even if the digitisation effort focuses on one small project it is very important to find out what quantity of specimens are going to be digitised. This number is the primary measure to consider how much time needs to be spent recording the information on the database yet is the aspect that most assumptions are made about when setting up a project. In fact, project estimates may be half the true number of stored specimens. This can cause many problems (not least a shortfall in available resource) which can make a project only partially successful or even fail completely. Even if this is the only thing you do to validate the specimens before the project starts, get to know how many specimens your project will digitise before the project begins.

What data quality?

It has been said that “quantity has a quality all of its own”. There is always a natural desire to produce as many records as possible in a database. Being able to give the number of records gives an easy metric which can be used to judge the success of a project. However, a simple listing of specimens may not be valuable to most users. Without suitable supporting data it is likely that significant subsequent work would have to be carried out to make the data useful.

Clearly, for each individual specimen, it is more efficient from an institutional perspective to completely process the first time it is digitised. However, there is the question of whether the funding body of a specific project would pay for the recording of information not directly linked to that project. Should the funding body refuse to pay, consideration should be given to matching resources to the shortfall in order to complete the work. If this is not practical, and with the limited funding available to most institutions it may well not be, recording the most commonly required data and the unique data specifically required by the specific project is a reasonable compromise between a complete data record and the limited needs of the current project. The mostly commonly required specimen data can be summarised as the accession number or barcode, collector, collection date, collection location, collection determination and the current determination.

Data capture or data interpretation?

The data recorded on the specimen is typically derived from the original collector’s notebooks and, as with any writing, can be full of errors. The question naturally arises of

whether the data should be captured as written (giving a historical perspective to the data), or corrected to give a more current interpretation (possibly correcting spelling errors or updating the country name to reflect political changes since the specimen was taken). Either way is an acceptable practise so long as it is recorded somewhere and the practise is consistently applied across the dataset.

One particular issue with original data is the area of taxonomic interpretation. Invalid names are frequently entered as a determination (a common example of this is citing the wrong author for a species name). In 2004, Meier & Dikow found that 62 – 73% of all determinations of *Euscelidia* were misidentified, so it is clear that this is by no means a small problem.

When recording the data this really leaves two options. From a historical perspective, this data should simply be left unchanged. This does make the data less useful for taxonomic research and so there is strong reason to correct the data. Doing this for every determination can be quite time consuming and so it is recommended that where possible a new determination is made with reference to a published source of names such as International Plants Names Index (IPNI).

Another potential aspect of digitisation is the addition of useful data not actually available on the specimen. Probably the most common example of this is the use of Geographical Information Systems (GIS) to provide location data. To do this specimens have to be geo-referenced (finding the place the collection was made and assigning it a latitude and longitude). On collections from the last 10 years it is usual to find latitude and longitude provided by GPS systems, but earlier collections rarely have this data, hence it must be added. This is a worthwhile endeavour but can be very time consuming as it requires additional research. If all the available collection location data is recorded it may be better to leave such value-added data to be added later. This can have potential advantages as it may allow a geo-referencing expert to concentrate on that one area.

Enhancing existing practices in the institution?

One important potential reason for digitising your specimens is to enhance the curatorial activities in the institution. Often this would require fairly minimal data from the specimen itself (often just the name), but significant additional data referring to the curatorial activities in the institution. A standard feature that is added to aid curatorial work is a unique barcode or accession number used to differentiate different specimens. This enables a user of the database to identify which particular specimen of *Quercus robur* it is out of a collection which may have multiple specimens that cannot otherwise be separated easily.

Imaging

In the vast majority of cases an image is a huge benefit, as it captures information that may not be easily recorded any other way. Occasionally imaging may not be appropriate; algae and bryophytes are particular examples where the value of imaging is debatable, as high levels of magnification are required to differentiate the characteristics of the specimen. Of course, it would be possible to take a high magnification image as well as the image of the specimen label, but this is an additional resource overhead. Imaging does have significant associated costs but these are typically outweighed by the benefits of having a good quality image. It is not the purpose of this paper to discuss imaging in detail as this has already been covered in a previous paper published by ENSCONET (Häuser et al, 2005).

Understand what digitisation will not do

Many digitisation projects fail to achieve their goals simply due to unrealistic expectations being placed on the database. So far, this paper has discussed examples of what a digitisation project can do for you, but it is also important to be clear about what it cannot achieve (McLeod and Winans, 1991). This section is not about what the databasing project will do, it is about ensuring that impossible goals are ruled out.

Databasing is not a money saving option.

Although certain activities can be made more efficient and less expensive, increased access to the data results in more queries to the institution. Introduction of information technology also has a commensurate cost in terms of computer equipment and maintenance (both of the machines and of the digital collection itself). Someone has to actually database the specimens, which requires a short term cost. Careful planning can allow the increased costs to be partially offset by the efficiency savings, but there is likely to be an increase in cost to match the increase in capacity created when you digitise your collection.

Digitising your collection will not create new information for you.

If the information is not already present on the specimen, it will take additional work to locate suitable references to create it. If the data are incorrectly written on the specimen, it will most likely be incorrectly entered into the database, even down to spelling errors in some cases. These deficiencies can prevent the system working as expected, causing the project to fail. Fortunately, databasing your specimens makes it much easier to identify these shortcomings in your data and will enable you to take preventative action. Exploring the uses of the resultant data by comparison with other records can add valuable additional data to that obtained directly from the specimen.

Specimens will still need to be physically stored and handled.

Although requests for individual specimens may fall due to the availability of the digital content, increases in data access will likely result in an increase in requests for specimens. No matter how detailed a specimen image is, there are still physical attributes of a specimen an image cannot hope to record.

When do you want the dataset be to be available?

Most projects are created in response to a perceived shortfall in the data already available, so it is not surprising if “yesterday” is your instinctive response to this question. In practise your goals may require a great deal of resources, particularly if you are working towards an institutional digitisation project. In many cases goals can be broken down into short, medium or long term phases. Chapman (2005a) breaks project goals into the following categories:

- **Short term.** Work that can be completed over a six to twelve month period.
- **Intermediate.** Data entry over approximately an 18 month period.
- **Long term.** Any project lasting longer than 18 months.

Given that many funded projects have attached deadlines it is often practical to map your digitisation goals to the deadlines of the projects at your institution. Naturally, if you are running a small project it is highly likely that your project deadlines are your longest term goal. For all projects though, defining practical short term goals is very useful as it allows

you to confirm you are making progress at an appropriate rate. This could be completing a specific subset of your research or supplying data to another institution (very useful for testing the practicality of your chosen method of data exchange).

Although an individual project usually has clearly specified endpoints, an institution usually has to define its goals for a longer period than can be funded by any one project. Setting goals and deadlines is still a useful activity as it helps ensure that new projects fit in with the institutions requirements.

Inevitably, whatever databasing work you do now has the potential to be useful in the future and ideally will be structured to reflect that – otherwise there will be a cost to the institution when it has to re-key the data for a specimen. However, it is difficult to know now what data will be important in 10 or 20 years time, so what do you choose to record? The most future-proof system would be to record everything, but this is also the most resource intensive option. As discussed under data quality, it may be more practical simply to record the most commonly required fields and accept that there will be a future requirement to record additional data.

Future requirements

Once a collection has been digitised, then there is a requirement to maintain the data over time. Quite simply, if the data are not maintained, there is a danger that its relevance and utility will decrease to the point of uselessness (Wheeler, 2004). For specimen collections, the primary example of this change is progress in the area of determining taxonomic names. To avoid this, the data should be maintained in the same way as any other collection. The data collection should be added to as the physical collection is added to, to maintain the relevance of the dataset in curatorial activities. Ideally, this would also include value added information such as geo-referencing data, so improving the usability of the data.

Your institution may have other databases holding additional information related to your specimens, such as seed, cultivation or genomic information, so a natural evolution is to integrate these datasets together. This will enable a better understanding of the total collection holdings of the institution and also enables researchers to access a much wider range of data than was previously available.

Technology is always moving forwards and integration of new techniques will always be an issue. A current example of this would be the use of mobile computing to record field collections. Trying to build new features like this onto a fully developed database is considerably more difficult compared to creating the facility when a database is set up. Considering the functionality you may wish to have in the future, even if they are not to be immediately implemented, will reduce any “growing pains” associated with extending your system in the future.

What are your current limitations and resources?

Having identified your aims, it is now time to consider the practical elements that you have available to deliver your project. It is entirely possible that insufficient resources are currently available to actually implement your desired project, but identifying the shortfall will enable you to decide how you will remedy the situation. This latter action is defined within the action plan. Many elements will be immediately obvious to the project planner, but are covered here for completeness. It may be that some of these limitations will actually

be advantages for your project or institution, but for most projects some additional resources will have to be committed to allow you to start your project.

Staffing

You will have several different roles that will have to be filled to help ensure your project will succeed. If a suitable person is available to your project it is a great help, but it is highly likely that some personnel will need specific training, even if it is just how to use the database and how to conform to the specimen handling requirements of your institution. This will naturally take some of the time allocated to your project. Assessing the degree to which you will need additional training will inform the level of time and resources required to complete your goals.

When thinking about staff and staffing costs, don't just think about the digitisers themselves. As with any other project, they will need managing, which takes time and hence has an associated cost. It is also good practise to include some data quality checking to ensure the work is being completed to the appropriate level required by your project. Depending on the size of your project, each of these roles may require a full person or may be incorporated into a single post. A good ratio would be one manager/data checker post per five staff or less. You may also need to consider including a database manager role. You must also decide who will handle the IT issues, even if it is just a case of buying a computer, installing the database and ensuring it all does not break down!

The typical roles associated with projects are:

Digitisers. There are several potential personnel pools you may be able to draw on to get your project done. It is quite likely that as project manager you will use one pool of staff to get the majority of the work done, but don't disregard the opportunity to use any chance you get to add data to your digitisation project.

Curatorial staff as part of their regular work. This is a very useful option if the database is primarily to be used to support curatorial work. It is also very practical when capturing information about loans as they enter or leave the institution. However, the curatorial team already have full time jobs to attend to, so the rate of data acquisition will inevitably be comparatively slow, if steady.

External contract staff/company. Passing data to an external team for digitisation can be a risky business. Despite the contactors promises there is no guarantee that the work is being done correctly until the results are returned to the hiring institution. This means that extra care has to be taken both in preparing the target collection before it is sent away for digitisation, and following digitisation when extensive data checks are required. Also, when dealing with external companies, making corrections to the data becomes more expensive, in terms of repeated shipping costs and the potential to be charged for the time required making changes to the specimen. It is rarely a free service, but is frequently cheaper than hiring project staff for the institution and has the added bonus of taking up no office space at the institution other than that of the project manager and data quality checker. However, the remote digitisers are rarely trained in handling natural history collections, so may not be able to interpret complicated data as well as staff in your own institution. If the data are recent and can be easily extracted from a printed file or label then this can be a valid way to go forwards.

Volunteer staff. Many institutions have volunteers who wish to get further involved with the ongoing work there. They can seem to be ideal candidates to get to do the digitisation work but this can be a double edged sword. Volunteers are usually eager to get involved but are doing so of their own free will and if they get bored will likely stop their involvement, perhaps without explanation. Sadly, while digitisation is very important, it cannot be described as the most exciting work to be under taken at any institution. Volunteers also expect to work to their own timetables and likely will not want to spend a full working week digitising specimens. Naturally, this means that volunteers will not achieve as much as full time project staff. They also will need office space for them to be able to do their work. For a small scale project, particularly one without great time pressure, volunteers can be a valid approach to digitising your

collection. Do be aware that there is an increased likelihood of staff turnover for your project, which may prevent you from making steady progress as you try to recruit and train new volunteers.

Visiting researchers. Depending on your institution, visiting researchers may be rare or common and they may want to look at any number of collections, making their own notes as they go. There are several issues here. The visitor may not want to look at the collection being worked on; they may not wish to enter the same quality of data that you expect and will almost certainly need it in the form they wish it for their own projects rather than yours. All of this tends to make researchers more suited to institution wide projects and will only be of very limited use to smaller projects.

Project staff. Making use of full time paid project staff is typically the best all round way to get your data entered at a good consistent quality, having the advantages of specially trained staff concentrating specifically on your project. It is also usually the most expensive and resource intensive option as you will have to pay more for trained staff (typically coming from a natural history background). Trained staff do have the advantage of being able to interpret the data themselves and be able to operate more independently than unskilled workers, thus requiring a lower level of management.

Students. At academic institutions, students may be available to do data entry. This can be a relatively inexpensive solution, especially if there are work-study programs involved that subsidise student work. Students may also find digitisation an entry level position that allows them to interact with, learn from, and become involved in the museum community. Using students may also make the project more worthwhile in the eyes of the institution and, in some cases, may earn release time for faculty to work on the project. In some cases, higher level students can be used for more advanced aspects of the project. For example, biology graduate students may be able to help with QC or specimen sorting or review prior to digitisation. Fine Art students may be useful in helping with the imaging process. Turn-over with students is an issue as is boredom and the potential distractions caused by student life.

Data owners. Researchers, other institutions and commercially available datasets can be a big help in providing ready made data, even if it is just the provision of standard values to populate lookup lists. Some care is necessary though, as Intellectual Property Rights (IPR) can become an issue, restricting the ways you are able to use the data.

Data experts. Curators and specialists in various fields will always be required for consultation purposes. It is almost inevitable that someone will be required to answer questions that arise during the project. Trying to divert someone who is not part of the project into helping you can a major hurdle, so getting them to agree to devote time to the project is a major asset.

Technical staff. Technical staff range from the IT support person who comes to setup your computer to the person who designed the database/spreadsheet you will be using. They all need some idea of the importance of your project and will probably have a role in maintaining the systems you use. If you are using a database system larger than Microsoft Access (and perhaps even then), one very important person will be the systems administrator for your database. He is the person who makes sure that the database is up and running, although he is not responsible for the actual data content.

Project management. There are two roles associated with project management. One is the project manager, who is responsible for the day to day running of the project and the second is the data manager, who is responsible for checking the data quality as well as maintaining the data. The data manager is also frequently the person who will continue to be responsible for the data once the project has finished. It may also be useful to have a project champion, a senior figure in the institution whose role is to support the project at an institutional level.

Data entry procedures

You must also consider how many can database at once. The total number of digitisers operating at any one time has several effects, notably the amount of physical space required for the project and the number of practical resources needed, such as the number of

computers. It also affects the level of complexity required of the database and potentially of the I.T. infrastructure required for the project.

Options include:

One person working at a single database. The easiest option but also the slowest in respect of project deadlines. For one person, you only need one desk, one computer and sufficient space to lay out the specimens to be digitised. Depending on the level of data security required there may be no requirement for any additional IT infrastructure. There is the risk, though, that if the individual computer fails, the entire dataset accumulated to that point will also be lost, likely causing the project to fail. Some form of data backup up is essential.

Several people using individual databases. As per the first option, this is a simple case of multiplying up the resource for one person, although the need to protect the data increases as does the likelihood that one or more of the computers will break down and need to be replaced. There is almost certainly an additional step required, which is to merge the separate databases together to form one single dataset at the end of the project. This means that using several people will reduce the time required to digitise a collection by a factor of the number of people involved (compared to a single person doing the work), plus the time required to set up the project and the time required to unify the data at the end of the project. The risks of losing large amounts of the data are somewhat mitigated by having the final dataset broken up into several parts. However, backing up the separate parts of the data is still recommended to prevent the need to re-record data. This will add an overhead to the project proportional to the number of people involved in digitisation.

Several people sharing the same database. This option combines the advantages of several people working on the same project, without the overhead of having to merge the data together at the end of the project. Protecting the data is also much easier as there is only one database to back up. In order to achieve this, some form of IT network will have to be in place, potentially adding to the resource burden of the project.

How big is your collection?

Knowing the total size of the collection will help you to judge the length of time it will take to digitise everything. An accurate assessment is important as it is easy to underestimate the actual size of a collection. It is also important to include a realistic assessment of the rate of acquisition of new specimens, as this must be included in your estimate.

Is access to your data restricted in any way?

As countries have become more aware of the potential wealth accruing from exploiting their native flora and fauna they have become much more protective of dissemination and use of the available information. This has resulted from and in agreements such as the Convention on Biological Diversity (CBD). These require permission to be sought for each country before specimens are collected. These Memoranda of Agreement (MOA) can have many different restrictions, including who may view the data. This may be restricted to one institution or even just one department in an institution. When sharing data, intellectual property rights (IPR) have to be respected and sometimes this requires institutions to sign legal agreements before data can be released. Due to the difficulties of agreeing legal agreements between countries, mostly due to lawyers refusing to sign legally binding agreements based on laws enacted in other countries, many IPR agreements are enacted as Memoranda of Understanding (MOU). No matter if the agreement is legally enforceable or not, these documents must be respected when you release the data. This is also why you should not harvest data from published websites without permission. The originating website may have had permission to publish the data but you may not have permission to use the data as you wish, which means you may be breaking the law. At the very least, you are abusing the trust of the publishing institution by not telling them how you are using large quantities of their data.

You may also wish to unilaterally conceal some of your own data. Rare species, such as those listed on the ICUN Red List, may be a target for commercial exploitation. Revealing data such as geographical location (especially as accurately as a GPS reference) can risk immeasurable damage to the wild populations. Hence you may choose to not to release this data, or only release a very broad definition of the location to protect the native population. There is some argument on this last point that the data are already available in other forms, such as digitised duplicates of the specimen or other collections from that location, but surely anything that makes the destruction or illegal exploitation of native species more difficult is a good thing. Although there is some progress towards a standard approach to sensitive data (see Chapter 6 of this *Manual*, which is based on Chapman and Grafton (2008)), currently there is no agreed-upon international standard, so you must follow the dictates of your conscience when implementing your project, in consultation with your colleagues.

Does your institution require you to use an existing system?

Should your institution already have a central database it is not unreasonable for the institution to expect you to add your data to that which others have already contributed. This can have the advantage of already having useful data for you to use and a ready made data entry system. However, it may restrict the way in which you can interpret and record the data. If this is the case, it may be better to use an individual system to record your project data, but it is vital that the data are provided to the institution at the end of your project in a format that can easily be imported into the central database. This means that any standard reference data requirements the institution requires should be followed in your database. In most cases, having a predefined database actually makes your project easier to deliver, so following institutional standards will normally be the best course.

Do you have legacy data (electronic or paper)?

Pre-existing research data may already exist in the institution, without being incorporated into a centralised database as described in the previous point. This data may enable a large volume of data to be established very quickly, although some time may be required to raise the data quality to an acceptable level or adapt it to your chosen system. Pre-existing paper systems are often easier to digitise (especially if typed) and are a strong candidate for external digitisation as little interpretation of the data is required. It is well worth taking the time with legacy data to check the recorded data against the physical specimens, in order to check both data accuracy and for any annotations that have been made on the specimen since the recorded data had been taken.

What are your physical requirements?

Digitisation does not take place in a vacuum. Staff and volunteers will be involved in your project will need a place to work. Things to consider include:

Where will the digitisation take place? There are three basic options for the location of the project work.

- **Digitise in the collection itself**, which has the advantage of working close to your specimens but typically there will be limited space to work, meaning only a few digitisers can work at any one time. If the collection is popular, the digitisers will be regularly interrupted with the consequence of a reduced work rate.
- **Establishing a dedicated area for digitisation** tends to remove the issues discussed above, at a cost in office space to the institution. However this does allow dedicated and uninterrupted work for a group

of digitisers. Moving the collection specimens to the digitisation area may prove a problem but careful planning can usually solve this.

- **Digitise in an entirely different location.** It is more difficult when the digitisation area is in a different building as moving specimens usually requires particularly careful handling. In this case, imaging the specimens in the collection and then using the image for digitising purposes can be a valid way of solving the transport problems.

All of these have been successfully used in the past to implement projects so long as proper attention is paid to the requirements of the chosen location.

Existing IT infrastructure. Proper attention to your technological requirements should be made. A full sized computer (PC or Macintosh) is better for daily work than a laptop and if already available will affect your choice of database. However, if you are going to be making collecting trips then a laptop is a practical requirement as you can then record observations in the field. Will you be imaging your specimens? If so, a camera or scanner will be required. If you wish to connect to the internet or anywhere beyond the confines of your desk some form of network will be necessary, even if it is just a telephone and modem. Consider what you have and what you will need and make a record of them.

Do you already have project deadlines?

If you have already started a project then you will already be working to deadlines that will affect your project. No doubt you will be very conscious of this limitation but recording the deadlines will help underline their importance to others, especially if you need their help to deliver the project.

Will you be working outside your institution?

Whether travelling to other institutions or collecting data in the field, working outside your regular place of business places special requirements on your project. Although you could record specimen data on paper and transcribe them later, it would be far more practical to make use of some form of mobile computing, such as a laptop or possibly a Personal Digital Assistant (PDA). If you wish to digitise specimens directly into your database, it will need to have one (or more) of the following pre-requisites:

be portable or

have a module that can handle data entry which is then imported into the main database or

has a web-based data entry system

The last option is rarely practical when considering fieldwork, as it presumes that you have a reasonable connection to the Internet no matter where you are. Even in today's world of powerful satellite communications it is rarely practical to have a long running connection when working outside of settlements (and sometimes not even then).

Funding

What funding do you already have, or which funding bodies could you realistically seek funding from?

Is there the will to change?

If there is no support for the digitisation work within the institution it will be much harder to implement any project. It is very useful to have a key person to act as a champion of your

project, and gathering supporters within the institution will also help. Without such support it will be difficult to run a successful large-scale digitisation project.

Produce a Tentative Business Case

Putting together a business case enables you to clearly set out your goals and limitations in a clear fashion. It should present to others why your project is worth undertaking. One of the most useful steps in getting a digitisation project up and running is to get the rest of your institution to accept and get involved with your project. This often means they will be willing to volunteer their time to look at elements of your project that cover their specialisations, enabling you to improve your data quality. If your curatorial team accepts the importance of your project, the task of getting hold of the appropriate specimens becomes much easier. Having staff willing to contribute resources to completing a project makes its successful completion much more likely and also increases its chances of attracting funding. You should also go beyond the simple statement of goals and limitations to answer practical questions such as:

What do you gain from doing this project?

Explain why it is worth doing for your institution and the rest of the world.

Is your project feasible?

It is important that you believe that your project is practical and achievable. Knowing you have set out all the arguments makes it much easier to do this. Many large scale projects are forced to answer 'no' to this question, in which case the goals outlined here should become inspirational targets that need to be broken down into smaller, short-term practical projects that attempt to meet the overall goals of the business case. In these cases, it is suggested that the business case is drawn up to cover the entire group of projects and a more detailed business case is set up for each smaller project.

Do your goals exceed your limitations?

Presuming your project is feasible, you are likely to have to estimate what additional resources you will need to complete your project. A few very lucky projects will have sufficient staff, time and equipment in advance to complete their project, but most will have to find additional resources from somewhere. Outline here what additional resources you believe you will need. Precise details of how this resource is supplied must be included in the action plan.

Rising above your limitations

Having considered your issues and decided that it is appropriate to proceed with your project you can finally outline a course of action, which will be filled out in detail in the action plan (or plans, if you choose to break a project down into several stages). Answering the following questions will help you supplement your existing resources.

Can changing working practices free up time to work on your project? Don't expect changes in working practises to provide a universal answer to all your issues. There is a reason why people work in the way they do and it is rare that simple changes in practises will release massive resources that can be repurposed to digitisation. However, it might be possible to gain sufficient time to allow a specialist to help with particular parts of your project. Changes in working practise can help in freeing up the needed physical space for a digitisation project. This is one area where gaining wide spread acceptance from the

rest of the organisation is a massive boon, as it is natural for staff to be distrustful of being forced to change successful practises in the face of new demands on their time.

Can other nearby institutions help out? It is entirely possible that you could arrange a joint digitisation effort with nearby institutions, using the resources of larger bodies to assist you (Snow, 2005), or sharing implementation costs with similarly sized institutions.

Who might fund your project? Many funding bodies exist, both on a national and international level. Bodies such as GBIF, the Andrew W. Mellon foundation or the Gordon and Betty Moore foundation all fund a number of projects every year. It may also be possible to gain funding from commercial bodies. It is important to be realistic about your funding opportunities as all funding bodies have many more applications for funding each year than they can actually fund. It is highly advisable to carefully research the requirements of funding bodies in order to shape your proposal accordingly.

Should your project be broken down into distinct units? If you believe your project as a whole is feasible it would not be surprising to say ‘no’ to this question. However, it is much easier to get a small project funded than a large one, so it is still sensible to give this question careful consideration. Again, exposing your business case to the rest of your organisation will help you quickly assess the viability of your ideas.

Can we do a proof of concept? Running a trial of any digitisation project is a very useful exercise. It will give you lots of practical information to help plan your project. It may not always be possible to do a proof of concept, but it is highly recommended.

The checklist in **Appendix A** should help you summarise your arguments and make writing the business case easier. Allow yourself time to reflect on the business case and seek constructive feedback from others, as it is always useful to question the underlying assumptions of any project. Some of the hardest feedback to take is from colleagues who may not initially support your project, yet in many ways this is the most useful. After all, it is easy to convince those who already believe it is the right thing to do, it is far better (and more satisfying) to convince doubters. Once you have updated the business case to take full account of the opinions of others, you can be reasonably sure you have a viable business case.

Pick a database solution

Now that you have a business case and the agreement of your colleagues, it is time to consider in detail how you intend to reach your goals. Although creating a business case may take considerable time, it is still important not to rush into the implementation phase. Poor decisions made at this stage of your project could still be with you years into the future, so a little time, patience and a willingness to test your decisions will be repaid by a smoothly running project when it finally begins. Do not worry if after implementing this section and section 6, you feel the need to come back and refine your ideas, this will result in a better project over the long term.

Selecting a suitable database is a complex decision which has to take account of many different factors. These are discussed in depth in the “Deciding on a particular database solution” and it is recommended the user reads that Section when making a decision about the database they will use.

Develop an Action Plan

Having written out a business case and selected a database to use, you will no doubt have many ideas about how you will implement your project. We should now try to prove those ideas will work when your project gets implemented. Simply start by writing down your project ideas in light of the issues raised in the business plan. Then check your ideas against the following points, altering your plans where necessary to take account of any issues raised.

Many of the issues discussed here interact with each other, so it is a practical precaution to go through the list at least twice and assure yourself that changes made as you work through the list still answer the issues raised earlier. Once you have completed this task successfully you can be confident that you have an achievable project that is as robust as you can make it.

Does your chosen solution match your goals, limitations, and resources?

If so, then you are in an excellent position to complete your project. For most though you will require additional resources to match the shortfall between your resources and your solution. Many of the questions here serve to detail that shortfall and to provide a baseline cost for your solution. Once you have that, you can either seek the funding or modify your solution to fit your available resources.

Will your solution handle your future requirements and what if it doesn't?

Ideally of course your solution will leave room to adapt to changing needs. Trying to design for all possible solutions can be very expensive in the short term (even if in the long term it is beneficial) and the costs may easily become prohibitive.

How many staff do you need?

This is somewhat dependant on how quickly you want your collection databased and is limited by the amount of physical space you have. Taking into account your deadlines and the size of the collection you want databased you need to get an idea for the number of staff you will use. It is better to overestimate the number you need rather than underestimate as over estimating making the project run more quickly while underestimating carries the risk of the project failing. A careful balance must be maintained here, if your project becomes too expensive then it will not get funded and your planning will be for nothing. The most effective way of judging the number of staff you need is to undertake a proof of concept, with staff actually digitising a significant number of species for real, using your planned workflow. Be very careful of using the estimated rates promoted by other institutions or database vendors, they are very often trying to make their institution/company look better than it actually is. As a very rough rule of thumb, assume a rate of 100 specimens a week per digitiser if you are going to database in great detail, with high resolution images. If you are not imaging, take a rate of 200 specimens a week and 300 specimens per week if you are neither digitising to a high level nor imaging. These figures are conservative, but you will be thanked for delivering or exceeding your project goals, which these figures should help you achieve. Don't forget to make allowances for staff holidays and absence from work due to illness, as these will affect the overall quantity of specimens you can digitise during the lifespan of your project.

How will you train your workers?

Training workers to use your system will take time and resources. Workers will need to know how to implement your workflow, how to use the computer and database software, and, most importantly, how to translate the specimens into digital information consistently, efficiently, and accurately. It is not enough to get data into the system; it has to be "good" data. This will only happen if the proper people are selected to do the work and that they are afforded proper training in how to do it.

How long will it take to build or implement?

All projects have some degree of lead time, often six months or more simply to find funding. However, buying and setting up computers, getting a working area and recruiting staff all take significant time. Outline what your lead time and add it to the duration of the practical phase of your project. For larger projects using professional staff, two months is the bare minimum time it will take to hire staff and will frequently be longer. This is broken down as follows: 1 week to get an advert written and published, 2 weeks for applications to arrive, sift and arrange interviews. Another week to hold interviews and finally a month for the successful candidate to work out any existing notice. Usually it is possible to get the physical requirements in place during this time, unless you also have to hire someone to complete those tasks, in which case another month is recommended. It is also wise to allow a little time at the end of the project to account for delays and complete any outstanding tasks (such as writing a project report). If you are going to include a website to show your data, the website development will run much more smoothly if it can take place following the main digitisation effort. If you do not do this, you risk unexpected developments in digitisation altering the website work, delaying it beyond the lifespan of your project. Website design can vary massively depending on the IT infrastructure of your institution, but to design a website from nothing, serving up large quantities of data will typically take around four months for one programmer.

What are you going to have to buy?

If you are recruiting new staff, you will most likely need new desks, chairs, computers, telephones and all the accoutrements needed for an office environment. You may need to buy a server to store the dataset on and a link to the outside world. If you are imaging, how are you going to image? Again, try to follow best practise for your specific specimen collection. Hopefully your institution will have much of this ready for you to use, but don't assume this is so before you start your project.

How much is it going to cost?

As you may begin to appreciate, running a digitisation project is not cheap. Make certain you include hidden costs to your institution and don't forget to budget for things like travel and subsistence when working abroad.

What will your workflow be?

In other words, what is the most effective method for your staff to proceed when digitising. The following factors will affect your workflow plan.

Number of digitisers. Larger numbers will allow you to specialise staff into particular roles (imager, databaser, geo-referencing and quality checker are potential examples). Be careful though as unless the specialisation is that digitiser's chosen vocation, it can cause boredom amongst your staff, which will slow down their rate of work.

Collecting and returning the specimens. This is a job that can typically be batched up so that a day or a week's worth of specimens can be collected at once, depending on the size of the specimens. If you are working in the collection you may even decide to collect each specimen as you need it. Consider how the curatorial staff will find a specimen you have taken away for digitising should they need it while you have it and agree this with the curatorial staff. Also consider if any preparation work must be done to prevent damage to the specimens from insects or poor handling, and factor time for that into your plans.

Specimen handling. It is also a good idea to document proper specimen handling procedures as your digitisers are unlikely to start as curatorial specialists trained in handling your specific collection.

How long will the specimens be absent from the collection? As the specimens may be in demand from other projects, minimising the length of time specimens will be unavailable is to be encouraged.

Location. If your staff is working in locations separate to your main collection, then the difficulty and time taken to transport the specimens to the digitisers must be considered. In this case, it may be better to transport digital images and database from the images.

Data Quality. Quality has to be considered in terms of the data you are going to record. For example, is it checked against recognised standards? Handwriting can be very difficult to decipher, adding a great deal of time to the digitisation process. Chapman (2005b) notes that it is far cheaper to capture data accurately than to correct errors later.

Adding value to the original data. Value can be added to digitised data in several ways. You can compare it to published standards, you can interpret the data to accepted lists (handwritten collectors almost always have to be interpreted this way as their signatures seem designed to confound the digitiser) or you can add valuable data such as latitude and longitude. Doing all of this takes time that should be allowed for, and the specific process recorded.

Imaging. If you are going to image the specimen, how are you going to include that in the workflow? Will it be before or after the specimen is recorded? Will a specialist do this work, or will the task be shared?

Data order. It may seem more efficient to follow the order of data as listed on a specimen label and enter it onto the database, but the order the data fields are presented in the database may not match that on the label and labels rarely have a completely consistent format (consider yourself lucky if yours do!). Experienced digitisers will develop their own preferred way of working, but when training give clear guidance on the most efficient way of entering your specimen data into your database.

Data checking. Always include time to check the quality of the data is up to your planned standards. It is very easy to spoil otherwise high quality work by simple data entry mistakes that can be quickly spotted by simply checking a random sample of data and checking in detail where needed. One way of doing this is to view the data in a table so it may be sorted so similar entries are grouped together, allowing mistakes to be more easily spotted. Redman (1996) noted that an error rate of up to 5% should be expected and it is far easier to correct those mistakes while it is still a simple matter to return to the original data (Chapman, 2005b).

Can procedures be overlapped? It may be that several stages of digitisation can proceed at once. This may be hard for an individual specimen, but working on several specimens at once may be possible. To take a simple example, a specimen could be being scanned while another is being databased, all controlled by the same digitiser.

What effect will staff absence have on your workflow? If you have a key member of staff on holiday or absent through illness, will this stop the rest of the digitisation procedure from happening? This may be unavoidable for a small team, but usually it is possible to make contingency arrangements to circumnavigate the problem.

Are there any bottlenecks in your plan? Careful planning will allow for a smooth process, but be careful to calculate the proper time allowance for each step in a process. Take this hypothetical example: It takes a digitiser 10 minutes to database a specimen, which is then handed over to be geo-referenced, taking another 3 minutes. If this was one digitiser to one georeferencer, there would be no delay in the process, and in fact the georeferencer would be under utilised. If there are four digitisers to each geo-referencing member of staff. This would mean that there is a bottleneck, as the georeferencer can only complete 3 full records before the next batch of four are ready, resulting in a georeferencing backlog of 2 records every 30 minutes. Considering how to handle that backlog in advance can save time when the project is up and running.

As can be seen, there are many things to keep track of when working out the practical process of digitisation. It is recommended that once you have worked out the process details you write it up as a user manual for the digitisation staff, as it will make the staff training much easier.

What happens when the project is over?

Once the digitisation has been completed the data still exists, so your planning should include an outline of what will happen to the data. You could use your database on another project,

possibly adding content to your existing dataset. Ensuring the data remains relevant is also important and ideally should be maintained following the end of the project. This could be done by having someone specifically responsible for updating the data or having the data widely accessible in the institution so it can be maintained by the curatorial staff in their day to day work. For every new database, this can take up to 0.3 full time equivalents (FTE) of a curators time (Snow, 2005).

Will your solution provide the appropriate level of data quality?

Decide what level of data quality you will accept. The use of published standards can greatly improve the quality of your data and can make data entry easier. Standards can be used for many areas as simple dropdown lists, but some caution must be observed, especially with older specimens as the terms used on the specimen may no longer agree with the current naming convention, countries are a particular example of this problem. In these cases, an ability to record the original name is very useful, but inevitably adds to the complexity of the database. Many standards exist and sometimes several may cover the same topic. Ultimately though, so long as one standard can be converted to another then it will not matter too much which one you choose.

Planning for the human element

Databasing is not a purely computerised system; it has a human element which must not be neglected. Digitisation requires training, not only in the process as described as above, but in the actual data entry system. Learning to interpret complicated data takes time and proper training. Poor training can be seen as the cause of a significant proportion of data error (Chapman, 2005b). Gaining experience in digitisation also takes time during which your staff will not be working at optimum efficiency, so allow for this when calculating the number of specimens which can realistically be digitised during your project.

How long will it take to database your collection?

There are two basic techniques to data entry. Detailed data entry means entering data carefully, using a maximum of lists, data checking and appropriate structuring of data to maximise accuracy and ease of retrieval. This produces the best data quality but at a high cost in time. Rapid data entry indicates the ability to enter data quickly and easily. However this implies a reduction in data checking and frequently less highly structured data. This can cause an increase in data entry errors and hence a lowering of data quality. Later correction of this data typically requires a high commitment of resources. You do not have to take up either of the two extreme options described above but achieving an acceptable balance between them is not easy and what may be acceptable for an individual project may not be suitable for an institutional system. Specimen quality can vary greatly and so it is difficult to set an average time required to database per specimen. This will also vary depending on the level of data quality that will be recorded and the level of accuracy that will be accepted. Once again, trials are the only practical way of developing realistic estimates of digitisation rates. Quality assurance (QA) is also a vital part of any databasing operation, allowing for error correction at an early stage, hopefully reducing the long term costs to an organisation.

Prioritise efforts

Undoubtedly the best possible result for any digitisation effort would be to digitise all the available data in the entire collection. However, this may take longer to complete and be too

resource intensive to do in any one project. In order to maximise the effectiveness of your resources you will have to prioritise your efforts, as discussed in the business case. To briefly recap, you can reduce the total number of specimens you will digitise by targeting particularly important specimens or a specific family or species. You could also reduce the amount of data captured by focussing on key data fields (typically basic collection information and determination information). Of course, you may need to combine both techniques to reach your goals or create a practical project. Here too, long term considerations will have an effect. If it is your intent to make the data available in the long term, then capturing a maximum amount of data per specimen will be more efficient for your institution than maximising the total number of specimens captured, however useful that target is in the short term.

Contingency planning/risk analysis

What happens if your project doesn't go as planned? By considering what might go wrong and preparing for it, you can help to ensure your project will run with a minimum of disruption whatever happens. Risk analysis is a project management tool and the reader is recommended to consult suitable project management texts to decide precisely how you will implement your risk plan. Here are a few simple things you will need to consider:

Staff loss or extended staff absence. How will you respond to the loss in digitisation rate? You may be able to hire new staff, or you may have to change to goals to cover the shortfall. It may be possible to extend the project slightly to allow others to complete the work. It may be politically difficult to ask for this time, but most funding bodies are sympathetic to unavoidable problems so long as the issue *was* unavoidable.

How will you address the issue of not making the required digitisation rate? This is a significant issue. Your project has a number of specimens it must digitise in a fixed length of time (if you don't have this, it is well worth setting a realistic target by which you can measure your performance). This implies a digitisation rate that must be set as a guideline. It will take time for the staff to be trained, so don't expect to hit the rate immediately, and if your targets are realistic, it will be exceeded at peak progress (allowing for normal staff absences). If they don't hit the target, then you should review your work processes to see if there are any avoidable delays, or consider getting additional staff involved. Paid overtime is a possibility if the project budget will run to it. Unpaid overtime is a sign of a badly run project as you clearly needed resources you did not factor into your project proposal. You may hit your targets, but your digitisers are less likely to wish to continue into another project, meaning you have lost a trained resource. If you cannot get more digitisers, reduce the number of specimens you will digitise until you have a realistic target.

What happens if your computer breaks down? At least, get a service agreement to replace your computer, but do be aware this can take time. Ideally, have funding to cover a backup computer, although this will rarely be practical on a small project. Also allow some time in the project to cover these delays.

Backup strategies. Protecting your data from hardware failure is highly recommended as part of any action plan. If your only computer fails and you cannot recover your dataset, your project has failed. This is not something your funding body is going to be sympathetic to, meaning it is unlikely you will be able to attract further funding in the future.

Malicious alteration of data. This is a rare occurrence, but is more likely if using an online system that could be externally hacked. Simple password security helps, which requires some form of database administration. Tracking basic information such as who last edited the data will also assist in the quality assurance process.

What will you do to document your solution/implementation?

Documenting what you do makes working during the project much easier, especially when training new staff. Proper documentation can make your data entry process more consistent among workers. It also enables you to look back and learn from your first project and apply that experience to your future projects. Nevertheless, creating documentation takes time and

can, itself, become a bottleneck, if the staff preparing it have other duties in the project. Appropriate planning in advance for your documentation can go a long way to making your workflow go smoothly and ensuring that your timeline is met.

Rating your project

Consider what indicators you will use to judge success or failure of your project. Don't simply consider how many specimens were completed in what time, but also include quality measures. Include a staff morale measure, as an effective project has the potential to yield trained staff for future projects, but if they no longer want to work for your institution you will have lost a valuable resource.

Is your solution a good return on your investment?

As stated at the beginning of this section, you should review your action plan several times to ensure all the various issues you need to resolve work in harmony during the actual implementation of the project. When this phase is completed, consider the project as whole and particularly consider the resource requirements. You must consider if the results will be worth the effort you will put in, as this will likely be one of the criteria a funding body will use. For a well designed project the answer should be a resounding 'yes'. If you are not certain, look again at the amount of work you are trying to do and try to focus more on the most important parts of the collection you intend to digitise, putting the rest off to a later project.

Eventually you will have a well planned project with a strong chance of being funded, allowing you to achieve your goals. When this is all written up, you should be able to run the project as soon as you secure your resources and/or funding. Now all there is to do is put it all together.

Running the project

Finally we have reached the point where you can actually implement your project. It probably seems like a lot of work has already been done without any visible results, but the reward is in a smoothly running project from the very beginning. During this period, there are still a number of things to do even before the project actually begins, but all are practical tasks designed to deliver your project.

Test your assumptions

If this is your first project, you will not know if your assumptions relating to the digitisation workflow and digitisation rates are valid. As discussed in earlier sections, practical proof is the only way to get this knowledge. Doing a small prototype project of maybe a hundred specimens (enough to gain some expertise with the chosen system) will help give valuable insights which can be used to refine your action plan. Having practical expertise with the system also enables you to better train your digitisation staff as you will be much more aware of the issues associated with digitisation. For this reason, it is highly recommended that whoever will manage your project staff undertakes this task. If you are building or altering a database as part of your project, this could form the first phase of data entry, in which case monitor the work very closely and be prepared to alter your workflow early in the project.

Should this alteration be left too late, it is entirely possible that the project will not recover the lost time.

Seek funding

All projects have a cost, whether it is met by the institution or by external bodies. A well developed project may take six to ten weeks to properly prepare (Snow, 2005). Finding suitable funding is something this paper cannot practically discuss in detail, as available bodies vary by country and by the exact nature of the collection being studied. Properly building up the business case and action plan can only enhance your chances of writing a successful project proposal, so hopefully getting funding will be straight forward if you have followed the suggestions in this section.

Build your database

Depending on the database you have selected, you may need to put time aside to have your database altered or even created. It is recommended that a modular design process is adopted (discussed further in later sections), enabling parts of the system to be tested, or even used practically, as the rest of the database is developed.

Hire your staff

Remember it may take a couple of months to get staff to arrive, so this is the first thing to do once you have secured your funding.

Develop documentation

Practical documentation such as a training manual and the design of the database are very important for future work. Suitable documents make it easier to develop experienced staff and maintain the database systems. Take time to develop good documentation before the project properly begins.

Arrange suitable office space

Your staff needs somewhere to work, so ensure it is ready for them when they start. Don't forget to allow sufficient space for any specialist equipment they may need above and beyond the normal office desks and chairs.

Buy equipment and set it up

Similarly to office space, this needs to be ready when your staff arrive as it is a waste of resources to have them waiting for the equipment to arrive.

Train your staff

No one arrives at an institution immediately ready to digitise. At the very least you will have to train them in your institution's procedures and probably in using the database. This is where the time previously invested in documenting your procedures will pay off, as it will enable your staff to begin working effectively much sooner.

Start digitising

At last, the main part of your project can begin.

Continually monitor the project

Be aware of the unexpected events, which, hopefully, you have prepared for by undertaking a risk analysis. Even having prepared by undertaking a prototype project, things can go wrong. Using short term goals and reviewing your achievements on a regular basis you can quickly react to unforeseen issues and adapt to overcome them.

Review the project

Once the project is over, whether it was successful or not, review it against the success criteria you laid out in the action plan. Consider what worked well for you and also what could be improved. This allows you to apply the lessons learnt from this project to the next one.

Once you have completed your first project, it is then time to consider what the next project will be. This should be easier having gained experience of the issues involved in a digitisation project, and you probably now have experienced staff and a database to make use of. It is still recommended to take the time to develop or update your business case and create a new action plan, as new techniques and opportunities will become available as time moves on.

Section 4: Organising Information and Representing Data

Which came first, the information or the data? For every definition of ‘information’ that uses the word data to define it there is one that defines ‘data’ using the word information. The discussions are as with the egg and the chicken - circular. Losee (1997) discusses some of these at length. Thankfully, it is not so much the definitions that are important in this context but that you have a clear understanding of three concepts:

You know things about your objects of interest, which we will call **object information**.

You will improve and extend the quality of your object information by the use of **reference and ancillary information**.

A computer can store and represent your object, reference, and ancillary information in certain ways. What it uses to do that we will call **data**.

Object Information (as you see it)

The information that you have access to, either already existing in digital format or waiting to be digitised, will fall into one of two categories: primary or secondary information. Primary information is assigned to identify an object or is taken directly from the object (i.e., from the label, tag, or field notes associated with a specimen). These are the things you know which should never change regardless of opinion (as long as it is correct in the first place).

Secondary information is used to describe or categorise the object or to associate the object with other types of information. Secondary information reflects what you know at a particular point in time and so will vary over time due to changes in opinion and increased knowledge.

Typical primary object information

Object identifier – The identifier that uniquely separates one single object out. This could be an accession number, barcode, LSID or any other method of assigning a unique value to a specimen and would be of the type *unique identifier*.

Collection event – Made up of collector, collection number and collection dates. These would respectively be of the type *person*, *identifier* and *date*.

Collection event location - A statement of the locality where the object was collected (*text* type). May or may not include coordinates (e.g. latitude (*GPS* type) and longitude (*GPS* type)), land ownership or management (*text* type), and elevation (*numeric* type).

Collection event method – A statement concerning how the object was collected.

Descriptive information about the object– The collector’s description of the living specimen (*text string* type). Information about the specimen preparation (*text string* type). Notes or remarks included with the specimen by the collector or the preparer (*text string* type).

Environmental – A description of the characteristics of the locality where the specimen was collected e.g. substrate (*text string* type), vegetation (*text string* type), associated species (*text string* type) & physical geography (*text string* type).

Donor information – The donor person (*person* type) and/or institution (*place* type), contact details (*place* type) and any special terms that might apply to the donation (*text string* type).

References – made up of title, date, journal name, collation, author

Typical secondary object information

Geographical (spatial) – Political geography e.g. country, province, district. Georeference(s) of the collection locality.

Taxonomic and nomenclatural - Includes the collector’s initial determination and any number of later determinations or revisions, along with the determiner’s name and date determined. Determinations are considered to be *taxon name* types, the determiner is a *person* type and the date is a *date* type.

Storage location - Which institution (*place* type) owns the specimen, its barcode or accession number (*unique identifier* type) and where it is stored (also a *place* type) within the institution. May contain a trail of various owners and storage locations over time.

Molecular – Although rarely recorded at the time of collection, it is increasingly common, and is good practice to associate DNA samples and sequences with voucher specimens. It would be considered to be of the *sequence* type.

Status markers and labels – This is a catch all for information about your object that relates it to other information of interest. Conservation status markers indicate that the object has some designation as rare, endangered, threatened, or sensitive typically through its name (i.e., it is a rare species). Type status designates the specimen as a type according to the rules of a particular nomenclature. Transaction markers can associate a specimen with one or more loans. Other markers may associate the specimen with projects, publications, mark it as part of some subset of the overall collection, designate its fitness for particular uses, or mark it for some particular consideration.

Remarks/comments – This is catch all for remarks about the object or about the information concerning an object that are not part of the object itself. Examples: “The jar seems to be leaking slightly,” “Not sure the ID is correct,” “The collector name was illegible, but I think this is an A. Smith collection,” “I can’t find this specimen in the collection – Felisa Jones 5/1/1999”.

Recording all of this potential information has a high short-term cost (Armstrong, 1992) but a long-term benefit is that the work never needs to be repeated. Partial recording of the primary information is cheaper in the short term but more expensive in the long term as, when the additional data is required, additional resources must be used to repeat parts of the digitisation workflow. In larger institutions this activity alone can take up a third to one half

of the total digitisation process, meaning the net digitisation cost per specimen is significantly increased.

Which of these types of information you handle within your system is determined by your purpose and the level of detail of each will vary accordingly. As a general rule try and make sure that the primary information is recorded as fully as possible and then be selective about the secondary data.

Reference and Ancillary Information

It is unlikely that your digitisation project will be confined exclusively to recording information directly associated with your objects of interest *and* that you will enter that information free-hand for each object. **Ancillary information** is all the electronic information that you will manage as part of your digitisation project that is not directly associated with the objects themselves. When ancillary information is used to constrain the values you can enter about your object it can be thought of as **reference information**.

Ancillary information can be as simple as the pair of values you will use to designate presence/absence of a feature, a list of values you might enter as labels for an object (e.g. male, female, hermaphrodite or sterile) or a much more complex set of information like all the fields associated with publications. Ancillary information often involves information about the reference information. For example, if the reference information is the name of institution where an object is held, the ancillary information gives you the name of the curator and her contact information. If the reference information is the Federal Status Designation for a specimen, the ancillary information could include the date that status was published in the Federal Register, its publication reference, and the populations to which it is applicable.

Typical Ancillary Information

- Nomenclature
- Geography
- Morphology
- Person Names
- Projects
- Institutions
- Publications
- Transactions

Data (as the computer sees it)

The important thing about data is that represents your information correctly and that you can get the same information out that you put it. Specimen information is usually stored in what we will call base units. These base data units are conceptual tools, which allow related data to be stored together and which can be joined in an unlimited numbers of ways to digitally represent your information. Later sections give examples of the different ways in which this is done in reality and your choice of database solution will take this into account.

These are some common data base units:

- Person – first name, last name, initials, title
- Taxon name – rank, epithet

Place – latitude, longitude, altitude, name, address

Date – day, month, year, century

Sequence

Chromosome number

Reference – Title, collation, year, Publication name

Basic data types

Knowing what data types best data is an important step in building or choosing a database solution. In essence there are three types of data: numbers, characters and dates. The choice of which can impact the way your database works on more levels than you might at first think:

Text/String fields are the simplest of data types and for the novice provide quick and simple data entry. What you see is generally what you get. They allow the user to type in both characters and digits, be aware though that although you are typing a number on the keyboard it will not be handled by the computer in the same way as in a true number field.

- **Pros:**
 - Raw data looks like it did when you entered it
 - Sorting produces expected results with letters
 - Easy to use in table format
 - Easy to format data for output
 - Copy and paste works as expected
- **Cons:**
 - Text fields take up more space than number fields
 - Text based lookup tables may be inefficient and slow if strings are long
 - You may have to type the same thing over and over again
 - Using the data for a purpose other than the one it was entered for usually involves reworking the data and may produce unexpected results
 - Sorting can produce unexpected results where numbers are involved or be impossible on very large fields

Parameters:

Length – this indicates how many characters can be accommodated in a field. Setting this to be too small will truncate your data. When importing data to a text field not all systems will tell you that this has happened and even cut and paste can be problematic. Setting it to be too long can have a serious impact on your ability to sort columns in some cases you cannot do this at all.

Padding – some text formats set a field size to a particular length regardless of the amount of actual data in the field. This can seriously inflate the size of your database so you should carefully check this is required.

Memo fields are a form of text field and are mentioned separately because they have some specific limitations. These types of field have no size restriction and therefore allow the user to be very verbose. They are especially good for description fields such as habitat. However, because of their potential size they are rarely sortable and difficult to search. They do not handle text formatting such as bold and italics at a character level and are erratic with regard to carriage return. They should be used sparingly and never for

data which is searched often.

Numeric fields are even simpler than text fields, allowing only numbers. However, they are often meaningless on their own and can be more restrictive than text fields. There are 2 types of numeric field: Floating point fields allow decimals, integer fields allow only whole numbers.

- **Pros:**
 - Number fields take up less room for storage
 - Allow you to use lookup tables efficiently
 - Easier to code form based data entry systems
 - Make it easier to atomise data
- **Cons:**
 - Usually require a text field as well to qualify them so you have to use 2 fields instead of one to handle your data.
 - Have to use lookup tables to handle text
 - Increased atomisation of data which makes exporting data more complicated
 - Data is not easily readable in table format

Parameters:

Length – determines the highest number you can enter into a field. It is usually expressed as the number of bits required to store the number.

Base – numbers do not need to be stored as Base 10, but may be in hex for example (Base 8). The computer and the programmer will love this but the user may find it difficult to interpret. Stick to Base 10 if you can.

Date/Time fields are modified number fields and should be used with some care. Most but not all store an integer which represents a particular date from a known start date which is 1, using this integer you then use the built in formatting to display that date in a variety of ways. However, different packages use different start dates, so you must check that the start dates are the same and that you are transferring the date and NOT the number that represents it when exporting or transferring your data.

Boolean fields are designed to reflect a dichotomous choice such as yes/no, true/false, present/absent, is/isn't. Typically, the underlying field is populated with either a one or a zero and entry is via a label such as the choices given above or via a check box, although in some solutions the actual entered value will be the same as the label. Care must be taken with Boolean fields with respect to default values and null values. For example, suppose you use a field "isSterile" to designate that a given herbarium specimen has no fruits or flowers. If no value is entered into the field, your particular solution may translate this as a zero, indicating that it isn't sterile. Alternatively, it may enter only a 1 or null in which case there is no way to distinguish not sterile from no value entered. The final possibility is that it will store a trichotomy of choices, 1, 0, and null (=nothing entered), although it may be more difficult to clear a field after a yes or no is entered.

BLOB (Binary Large Object) or Container fields allow you to store files such as images, sounds, videos, documents and other binary data within your database. You may have to define in advance the largest file size you can store in a given BLOB field. Of course, storing files within your database can dramatically increase the size of your database file which can affect performance. Some solutions allow links to be stored instead of the files themselves. In this case, the database file doesn't grow so big, but moving either the linked file or the database may cause the link to be broken.

Calculated fields

Often you will want to handle information which is actually calculated from fields which you have measured and have already been entered into your database. As always there are many ways to do this but the things to bear in mind are:

1. how often the formula and its parameters are going to change.
2. how often the measurements that are used to calculate it will change and
3. the final use of the results,

If the formulae and parameters are relatively stable then it may be worth considering creating fields in your database which store the results of a calculation rather than hard-coding (typing in by hand).

If the measurements are subject to change then you need to consider how you are going to keep track of the changes in the database. Often a linking table will help you out.

If the results are going to be widely used and add value to the original data then they are probably worth storing in your database for others to use. If not then you may be as well to do the calculations in a familiar environment and not keep them in the database or do the calculation as part of your final display. It is not wrong to use packages other than your database for calculations.

Functions

Calculating fields and using formulae in your database will bring you into contact with functions and this is where your data type becomes important. In most cases the functions that you can use on a text field cannot be used on a number field and vice versa.

Remember that although Date/Time fields appear to behave as text they are really numbers underneath which is why you can add and subtract them. Using functions to write equations and formulae can significantly reduce data entry time but require some investment in terms of coding and working out the logic.

In general if you can code it then do so, it will mean that all entries are always calculated in the same way and you only have to do it once. Of course if you don't have the programming expertise to hand or the time to learn a new language then re-use someone else's code (that's how the real programmers do it). If all else fails enter it by hand at least you've got it.

Special characters and encoding

Special characters include diacriticals, accents, mathematical symbols and non-latin letters. Where data is entered from different original languages these are important, however, inconsistently used they can be confusing. For example are Wurdack and Würdack the same person? If you are entering data from scratch decide *a priori* whether your data entry staff will use them or not and stick to it.

Encoding is how programmers map special characters. There are different methods and you will need to know which your database solution uses especially if you need to import legacy data.

Both topics are discussed here:

<http://www.nada.kth.se/i18n/iab-charsets/terminology.html>

Data Views and Storage Formats

Most database solutions use a combination of four ways to present your data to you. The data entry view, the storage format, export formats and multimedia displays. Any of these may be presented to the user as rows, i.e. a table or as a screen of data entry boxes i.e. a form.

Explaining the difference between these things to the users is a useful exercise and will avoid much confusion especially where your data entry personnel are also responsible for the way the data are presented to other audiences but are not programmers themselves.

The data entry view

The data entry view is the place where data is input to your database. Introducing a new data entry system where one exists already (particularly ones created by the users themselves) is often the make or break of a digitisation project. It is not impossible but getting these individuals on board early in development is crucial. There are often very good reasons that an existing system behaves in the 'quirky' way that it does! Don't be dismissive be prepared to listen to the logic they may know something you missed. Only then explain your more elegant solution, twice if necessary. As long it's sensible and doesn't slow them down too much they'll come around. Of course where nothing exists you will need to assess your data-entry staff and select something that maximises efficiency but does not compromise correctness.

The more structured and atomised the underlying database structure the more likely it is that the data entry view will be a form-based solution. Data entry is often slower in these systems but data checking can be very stringent. To increase speed of entry some systems use Rapid Data Entry (RDE) tables which can then be imported but these have less stringent data checking. Which of these you opt for depends on who is doing your data entry.

It is important to realise and/or explain that this view is not really meant for displaying/presenting data and it should be streamlined for data entry. If you are specifying a data entry system you should bear this in mind. Don't try to over-engineer it just to make things look nice. It is more important to be flexible and to allow for as many types of user as possible to get data into the correct parts of your underlying database. You'll never get it a single data entry solution that is right for everyone.

For example: It may be the case that your database is held in SQLserver and have a MSAccess front-end for routine data-entry within the institute and a web form for outside data-entry.

The storage format

The storage format is how the data are stored in the database. If you have a suitable data entry view users will not need to be concerned with this format. However, in general, the more complex the data model the more likely the storage format will make use of link tables, record IDs and objects. In these cases very little can be gleaned from seeing the actual data. Of course for a simple flat data model this format may be the same as the data entry view.

Export formats

These are many and where data will need to be analysed in packages outside your database they can be a crucial criterion in choosing your solution. In these cases the database becomes a repository for primary data rather than a work area.

Multimedia displays

This is where data should be manipulated so as to target different audiences and highlight specific things. It is also the area most often confused with data entry. If your data have been entered well and in the appropriate formats you can display them pretty much as you please so it is important to distinguish clearly between how and where you enter data and what you display.

It is often the case that this part of the digitisation project is left until last with the least resource allocated and time allowed. It is the way an external audience judges your project and is as important as the underlying database itself. If resources are short it is worth considering this as a separate project with its own line of funding and staffing.

Standards

What is a standard?

A standard is a document approved by a recognized body that provides for common and repeated use, rules, guidelines or characteristics for products or related processes and production methods. One of the most common misconceptions is that there is a collections database standard somewhere which will tell you how to build your collections database or information management system. It does not exist. There are many different collections databases and other information management systems in use and they are not underlain by a common, standard data model.

Standards, however, do exist that can affect biodiversity informatics activities, including the design of collections databases and information management systems. For our purposes here, these standards can be grouped into four broad categories:

Data Exchange Standards. These standards, also known as transfer or transport protocols, are used to organise and format information so that it can be exchanged or combined regardless of source. The most commonly known data exchange standards for collections data are the Herbarium Information System and Protocols for Interchange of Data (HISPID) (Conn 2000), Access to Biological Collections Data (ABCD) (<http://www.tdwg.org/activities/abcd/>) and the Darwin Core (DwC) (<http://www.tdwg.org/activities/darwincore/>). Exchange standards give the headings, fields, tags, or elements with which to organise your data. ABCD and DwC are both expressed as XML schemas. ABCD has a hierarchical structure and is intended as to be a comprehensive and detailed format to model biological collection information. DwC has a much simpler format and is designed to facilitate the exchange of the “most important” information that might be generally useful.

Standard datasets are used to create “controlled vocabularies” for certain kinds of information. These can be extremely useful when used as the basis for lookup lists and reference tables. An example of a standard dataset is Brummitt & Powell’s Authors of Plant Names (1992) which is recognized by the International Convention of Botanical Nomenclature as the standard for author abbreviations in plant names. Another is ISO 3166 which is a geographic standard for coding the names of countries and their principal subdivisions (e.g. states and counties/provinces). These codes can be very useful in constraining the values for your geographic administrative units. Saying a dataset is a “standard” doesn’t mean that it is necessarily the only choice available for a certain type of information. For example, Federal Information Processing Standard (FIPS) 10-4 has an alternative listing for two letter codes for countries which is different from that of ISO 3166. Saying a dataset is “standard” also does not mean it will fit your needs perfectly or even well. For example, neither FIPS 10-4 nor ISO 3166 has an entry for England, Wales or Scotland which might be a problem if you used these standards for your drop-down list for country as it appears on your herbarium label.

Best Practice Documents are guidelines to help standardise methodology and practices and are generally vetted by an organisation or society. Examples include the American Society of Mammalogists’ Documentation Standards for Automatic Data Processing in Mammalogy (McLaren et al. 1996) and the

documents produced for the Global Biodiversity Information Facility by Arthur Chapman (Chapman 2005a, 2005b, 2008).

Technical Standards is a catch-all term for standards that do not fit in the previous categories. Typically, technical standards affect the design and implementation of systems that allow the exchange, presentation, and manipulation of data. Software developers use technical standards to build support for the interfaces and encoding into their products and services. For example, the TDWG Access Protocol for Information Retrieval (TAPIR) specifies how to use HTML to transfer XML-based request and responses to access structured data (i.e. data in ABCD or DwC format) stored on any number and type of distributed databases. Another example is the OpenGIS Web Map Service (WMS) Implementation Specification which supports the creation and display of map-like views of information that come from multiple sources.

Standards provide the common language, rules and protocols for the sharing and interpretation of information (Conn 2003). Understanding and using standards can increase the quality of your information system, streamline development, and increase interoperability of your system and information with other systems and information. On the other hand, there are many standards and it may take a high level of expertise to be aware of standards that may be applicable to your situation and to choose which standards are best for your purposes.

Standards Bodies

These are organisations which both develop and maintain standards. Increasingly they are cross-fertilising and looking for ways to link standards together in meaningful ways. There are many international standards bodies and even more that operate at a regional or national scope.

Bodies

Biodiversity Information Standards (TDWG):

<http://www.tdwg.org/>

OpenGIS Consortium (OGC):

<http://www.opengeospatial.org/>

International Standards Organisation (ISO):

<http://www.iso.org/iso/en/ISOOnline.frontpage>

Lists

<http://www.consortiuminfo.org/>

<http://bubl.ac.uk/link/i/internationalstandards.htm>

In addition, there are non-standard bodies that have online resources and hold meetings and workshops that serve as useful starting places to help understand standards and their role in our community:

Global Biodiversity Information Facility (GBIF):

<http://www.gbif.org>

The Society for the Preservation of Natural History Collections (SPNHC):

<http://www.sphnc.org>

Natural Science Collections Alliance:

<http://www.nscalliance.org>

And the many taxonomic societies.

Data Quality

What is data quality?

The key to remember is that it is all relative. The terms “*fitness for use*” (Chrisman 1983), “*potential value*” (Dalcin 2004) and “*defect-free*” Redman (2001) have all been used to describe data quality and indeed all of these should be considered as indicators of whether your data is any good or not. In the end though it all boils down to whether you can use your data to do what you want to, whether you can explain what you have to others and whether it can be used by someone else for something completely different.

Chapman (2005a) states that data quality should play a role at every stage of the digitisation process and this is crucial as it will allow you to prevent problems arising in new data and correct things in existing data. A simple way to assess the quality of both the data that you have and the data that you aim to create is to use Redman’s (2001) list to think about its:

- accessibility;
- accuracy;
- completeness;
- consistency with other sources;
- relevancy;
- comprehensiveness;
- level of detail and
- ease of interpretation.

These qualities are relevant regardless of the size of your digitisation project and so it is important to decide how, given your goals, you are going to address each. It may well be that you have to prioritise them given your working limitations, but they should be accounted for in your action plan.

Entering new data

Of course if you are starting from scratch you only have to work out what data you need to have in order to get the results you want in the time you have, but the purpose of your project is still the most important thing to know and to document.

One way to help create high quality new data is to use lookups, dropdowns and/or controlled vocabularies. These are lists of standardised data/terms from which one or more options may be selected for a particular field in a database. As the data values have already been checked, the use of these lists has the advantage that data accuracy is improved, although it does not remove the possibility that the wrong option is chosen. Hierarchical lookups which filter themselves is another useful way to make entry more accurate. If you can use lookups then do, it will reduce the time spent error checking. There are, however, limitations and drawbacks to using standardized datasets in look-ups that should be considered at the outset of your project. First, they must be obtained, formatted, and perhaps augmented before data entry starts. This may impact your ability to start digitisation in a timely fashion. Secondly, these standard datasets may change or be updated after you have imported it into your system. It is not necessarily a straightforward process incorporating and reconciling these changes with the rest of your existing data.

An alternative is to use lookup lists in a less strict fashion, still allowing data entry for trusted users and restricting others.

The use of anything but simple lookups will, in the majority of database solutions, increase the programming overheads and may complicate the database structure itself.

Importing existing data

In the majority of cases you will have existing digital datasets (legacy data) which you wish to re-purpose, incorporate, merge or build upon to help you to achieve your goals. In other cases, your workflow may entail creation of datasets outside your system and subsequent ingestion. Whether you intend to move these data to a new system or add functionality to an existing database system, the first and most important thing you need to do is work out what you have and why it was created. It is rare that a dataset can be transferred from one system to another without some work being done to it. The first principle of practical data quality assessment is ‘purpose.’ Once you know what you have and how combining your datasets benefits your aims for this digitisation project you can decide what extra information you will need to achieve your goals.

Remember that legacy systems have rules all of their own and just because data is held in database software does not mean that it functions as a database. Time spent assessing each table and field to determine both its purpose and actual content will save you time in the long run. You will inevitably have to ‘clean-up’ legacy data and the time taken to do this task should not be underestimated.

Maletic and Marcus (2000) define data cleaning as:

- Define and determine error types
- Search and identify error instances
- Correct the errors
- Document error instances and error types
- Modify data entry procedures to reduce incidence of similar errors in future.

This is discussed in more depth in Chapter 4 of this *Manual*, which is based on “*Principles and Methods of Data Cleaning*” (Chapman, 2005b), but some things to look out for are:

Field names

Field names can be misleading in many ways.

Case 1: different disciplines use the same terms to describe completely different concepts. The term ‘valid’ in a zoological names dataset does NOT mean the same as in a botanical one. So despite being used correctly in both cases they are not equivalent.

Case 2: the contents of a field may bear no relation to the field name at all. This may have occurred because there was nowhere else in the system to enter the required data or because the user did not understand the field name.

Case 3: a field may change its use. A field labelled ‘date’ may have originally been intended to hold the collection date for a specimen but a second user thought it was the determination date. So, while the data itself looks fine there are actually two different bits of information.

Column/field order

This should not matter in a well-designed database; however, not all databases are well designed.

Example: There are 2 fields in a spreadsheet one to record determination date and one to record type validation date. You know what they are because the field to the left says Det by and Validated by respectively. However, they are both labelled 'date'.

Because spreadsheets use cell references to identify data elements and columns this is perfectly valid. However, you will experience problems when you import them to a database package because the field names will conflict. At best it will prompt you and give the option to rename at worst it give the field an auto-name. It is better to give the columns distinctive names prior to import.

Rows vs. records

In a database table, rows represent records and each record represents a unique instance of something (e.g. specimens, people, publications, etc.). Each record is comprised of number of fields which exist whether or not they are populated with data. Each record has the same data or potential data. Data that come in from text documents or spreadsheets are not necessarily organized this way. The data may be hierarchically organized with headers that are repeated only once for each instance.

Case 1: Rows in a database table have been used to hold dividing label information. In the table below records one and five represent headers in the original document that have now been parsed erroneously over the first several fields in the database.

ID	Barcode	Collector	Coll num	Location	Name	Det date	Country	Coll date	User
1	The	Adam	Smith	Collection	1960-9				
2	98987	Smith,A.	90	BM	S. aph.	2/6/1969	Ecuador	Mar 1969	yy
3	98988	Blogg,B	1	KEW	B. perr.	5/12/2006	UK	Jun 1908	xxx
4	98989	Anon.	306 (I think this is Smith, A)	NY	E. sup.	8/8/1971	France	Sep 1701	xxx
5	The	Richard	Spruce	Collection					
6	10001	Spruce,R.	5040	BM	M.aus.	1/1855	Ecuador	12/1852	Yy
7	10002	Spruce,R.	5041	BM	M.apr.	1/1855	Ecuador	10/1851	yy

In a text document or spreadsheet, information may be repeated inconsistently from one record to the next.

Case 2: The row for record 4 had a comment inserted after the collection number.

Formatting and data types

In a database, data in a given field all have the same data type. This is not the case when importing from a text document or spreadsheet. For example, a date column in the original document might be mostly populated with data in a mm/dd/yyyy format, but occasionally have a cell with a value like “1/11-13/2001” or “Spring 196?”. These will not come into your database properly. Even worse, something like “May 2001” may represent formatting on an underlying date of 05/01/2001. This will import correctly, but does not accurately reflect the true collection date information.

Never assume that the contents of a field or similar fields are formatted consistently. Unless they were entered from an un-editable lookup list they won't be and even if there is a lookup list it's not guaranteed. Do NOT underestimate how long it will take to reformat all the values in a dataset that need it. Automatic parsing scripts will only get you so far, be pragmatic about how long it takes to develop them, at some point you will have to do some by hand.

Combining datasets

You may have more than one dataset with different original purposes in different systems and even different formats. While it is perfectly valid and often desirable to combine them you MUST be aware that with different original focuses the limitations of one dataset may negate the usefulness of the other. Although the quality of an entire combined dataset is not as low as the worst single dataset contained within it, be aware that it is not as high as the best. Data quality may in fact be better preserved by linking datasets together rather than merging them and it may also be the case that merging is not as simple as first impressions suggest.

Example 1:

Dataset 1: is a specimen dataset created by population ecologists to look for genetic drift in a particular species complex.

Dataset 2: is an accessions register.

A field in the 1st dataset is called ‘Sequence?’ and uses a ‘y’ to indicate that there is a DNA sequence to go with this voucher specimen. Null values indicate that there is not.

The 2nd dataset does not have this field at all.

If you combined to two datasets without altering the 1st one to use ‘n’ to explicitly state that a sequence does not exist, a null value could have one of two possible meanings and the user would not be able to tell. The quality of this field is now reduced.

Example 2:

Dataset 1: A taxonomic treatment of *Corallina*

Dataset 2: A Coralline algae type catalogue

The 1st dataset records information about a particular taxon, and also the specimen which has been designated as its type. A field called ‘taxon’ is used to store the basionym of the taxa in question.

The 2nd dataset records all the specimens in the herbarium which are currently in a type folder and uses a field called ‘taxon’ to store the current name of the specimen.

Here the same field name has, quite correctly within the context of the table itself, been used to record two very different pieces of information and the rows are not analogous. Merging of these datasets while possible would not be straightforward.

Language

The default language used in a database should be appropriate to the database entry staff and to the primary users. It may be that more than one language may have to be catered for in the database. Apart from the use of common terms in a common language such as Latin, this is a huge complication. Not only do the data have to be recorded separately in each language, suitable procedures have to be put in place to display the appropriate version of the data. Maintenance of the data is also made more complicated, as two versions of the record must be updated. Automatic translation of the data can be attempted, but these are not always accurate and hence reduce the reliability of your data.

Maintaining multiple languages in a database can be done but has maintenance issues which, may out weigh the data–entry usefulness. At the very least, the use of proper encoding is imperative if any of the languages are non-Latin.

Intellectual Property Rights

The issue of intellectual property is vast, complex and outside the scope of this section.

However, it is an issue that every digitisation project should be aware of, and which, to as great an extent as possible, it should address. IPR will impact your project both in terms of using information and data to create the digital resource, and also in how that resource is disseminated to the target audience. Be aware that even though you and/or your institute may have access to and regularly use a dataset for research you may not actually have the right to publish it in its native form (either on the internet or on paper) for a purpose other than that for which it was originally given.

As a general rule always try to find out what the original source of the dataset is and document what you did to the best of your knowledge, always get permission to use it and always acknowledge the source. Check your institute's guidelines and any local legislation, as rules differ from country to country.

These are places to start:

GBIF

<http://www.gbif.org/News/NEWS1174645079>

USA

<http://usinfo.state.gov/products/pubs/intelprp/index.htm>

UK

http://customs.hmrc.gov.uk/channelsPortalWebApp/channelsPortalWebApp.portal?_nfpb=true&_pageLabel=pageLibrary_ShowContent&id=HMCE_CL_000244&propertyType=document

Section 5: The Data Model

Introduction

If you read the last section, you have perhaps started the process of identifying the primary and secondary object information that you are interested in recording and the ancillary information you are interested in maintaining as part of your digitisation project. You have

some idea of how the computer is going to see your data and how you want your system to present the data for different uses. Now you need to address how that information is going to be organized and handled as data by a computer system. This is the topic of the data model.

We will start our discussion of the data model with respect to a simple case, the catalogue, and introduce the first organising concept, the **base unit** of interest. We will then move on to a more complicated situation in which both the base unit and the **focus** of the data model becomes important considerations in the model design. These points will allow us to better understand the next subject which is the more complex, modular design of information management systems.

The data model is the foundation for an implementation in which the data are entered, viewed, manipulated and made available to others. We discuss some basic concepts in data model implementation including the structure of the implementation system and the importance of distinguishing between what the model can do and what the implementation itself can do to make your data usable. We then address some key concepts involved in how the complexity of the data model impacts the way information is handled as data in an implemented information management system.

The Base Unit of a Simple Data Model

Let's start by considering the most simple data model that you might realistically use as you digitise a collection: a catalogue. A **catalogue** is a representation of a collection of objects as a list. A catalogue differs from a simple list in that it contains descriptive information about the objects in the list. To exemplify the role of the base unit, let us do a thought experiment in which the objects of interest are the arthropods in an invertebrate collection. Your goal, as curator of this collection, is simply to generate a catalogue of your holdings so that you and others will know what you have. Seems simple and straightforward so far right? Even in this simple example, however, there are a number of different base units you might choose and different ways of organising your data model.

The individual as a base unit

In the first case, we will choose the individual pinned insect as the base unit. To have your electronic catalogue be an accurate representation of your collection, each record and each insect *must* have a unique identifier. These identifiers fulfil two functions. First, they match each record in your database uniquely with each specimen and, secondly, they identify each specimen as distinct from all others. Note that in this data model all of the information associated with the specimen except the identifier can be considered descriptive and, to some degree optional with respect to the data model. A specimen can be catalogued, for example, so long as it has a catalogue number even if it doesn't have a taxonomic name associated with it.

The base unit in degenerate data models

In the second case, the collection consists mainly of pinned insects, but you also have collections of spiders in vials of alcohol and your thrips and other small invertebrates are on slides. The vials and slides can have hundreds or thousands of individuals in or on them and it is impractical if not impossible to assign an identifier to each individual. Instead you will assign unique identifiers to each pin, vial or slide. Your logical base unit now is not the individual, but the "preparations" in your collection be they pins, slides or vials. This data model is **degenerate** in the sense that each of your base unit preparations may in fact

represent a collection of objects of interest. The degenerate model has implications for the information you record and its interpretation. The taxonomic name associated with a slide, for example, may represent the identity of each individual on the slide or it may be at a higher rank that reflects the lowest rank that all the individuals have in common (i.e. they are all members of the same family). You might have aggregate or summary fields such as the count of individuals or a list of the sexes and developmental stages represented in the vial.

Of course, the greatest conflict in a degenerate model is when you want to associate information with an individual or a subset of individual within your collective base unit. For example, you wish to publish a type based explicitly on only two of the spiders in a vial or you do DNA sampling on only one fish in a jar of fishes, but how do you record this information in your database and how do you represent these events with your preparations? The answer is not simple. It may require use of note fields or linked tables. You may need to use subunit markers like gill tags or it may be necessary to promote a subunit. So for example, take out the two spiders and put them in a separate vial with a new unique identifier.

Suppose, your collection is very large and your goal is simply to know what you have as fast as possible. You do not have tags on each of your specimens and for your purposes it would take too much time and effort to label each specimen. In this case, you might be tempted to create a catalogue based on taxonomic name. Your catalogue might contain a list of unique names as your base unit with the count of specimens or preparations for each name as the main descriptive information you record. Another alternative would be to use the drawer as your base unit. Your collection consists of a number of drawers in a number of cabinets, so why not just record each drawer, its cabinet location and the taxonomic identity of the specimens in each drawer and perhaps the number of specimens or preparations in each drawer? If it fits your purposes, this may be the way to go.

The problem with degenerate models is that, by definition, they exclude information about potential subunits. It not necessarily “wrong” to use a degenerate model if the information is unnecessary for your purposes,. Keep in mind, however, that a degenerate model may not easily be convertible to a model based on another base unit if, in the future, you decide your existing solution does not meet your needs. For example, if you developed a system based on drawer location and counts per drawer, converting to an individual or preparation based system might take virtually the same time and resources as starting the new system from scratch.

The Focus of a Data Model

In data models more complex than catalogues, one must consider both the base units employed in the model and the focus of the model. Let us use the example from the last section of a collection of pinned specimens in an arthropod collection. In a specimen based model, the base unit is the specimen (or preparation) and it is also the focus of the database. Other information such as the taxonomic name and collection information are added as attributes of the base unit specimen. But we know that, particularly for arthropod collections, when one organism is collected a large number of other organisms are collected at the same time in the same collecting event. So perhaps it would be more efficient to have the focus of our data model be the collection event itself. In this case, we give each collection event a unique identifier and associate a lot of information with that event like who was there and when and where it happened and why. We could then associate the specimens, the “what was collected” as conceptual attributes of the collection event.

The difference between these two models is profound. For example, in the first case, each of the specimens has a unique identifier, but when the focus is the collection event, this need not be the case. The collection event could be linked to a number of uniquely identified specimens or it could be linked to degenerate specimen information (1,200 thrips, 200 cockroaches, etc.) or it could possess some combination of a descriptive list of what was collected and individually identified specimens.

Note that in our simple case of a specimen based catalogue of arthropods in a collection, we still are likely to enter information about the collection event and vice-versa for the collection event based system. The difference is how that information is converted into data in the data model. In the first case, the collection event information might be entered free-hand for each specimen or it could be cut and pasted from one record to the next if they were collected during the same event. But in either case we do not have sufficient information in the system to tell us much about a given collection event. Doing a grouping on the collection event information doesn't help much either, since the information might have been entered somewhat differently for each specimen record. On the other hand, since our focus in this case is the specimen, we don't have the burden of having to enter much about the collection event in order to populate our specimen records.

In terms of data model, then, the next step more complicated than a simple catalogue is a system with an identified base unit, a single focus, and which contains additional descriptive information about this base unit. This descriptive information may include one or more lists, but these lists are treated as multiple-value attributes of the focus information. The focus is used to make data entry efficient and directed towards the purposes of your digitisation project. The focus dictates the priority you will give to recording and maintaining information in your system. The focus also determines what kind of output you are likely to generate from your system.

The following are some examples of common information products and purposes in our discipline and their relationship to the information they use and the foci of the information systems by which they are generated.

Floras and Faunas

These are **taxon based** products. The focus is the taxa within some superset of higher ranked taxonomy and implied geographical extent (e.g. the flora of North America or the mammals of Queensland):

Taxonomic and nomenclatural information, providing the currently accepted taxon names, synonyms and discussion of the application of names to members of this fauna or flora.

Geographical (spatial) information, providing the distribution information for each taxon.

Publishing information, providing the reference information concerning each taxon or each taxon name.

Descriptive Data for each taxon.

Remarks and comments of the author or cited authorities.

Images representative or illustrative of each taxon.

Voucher Information including specimens that were observed by the author, collected by the author in support of the flora or fauna or which otherwise document a taxon's inclusion in this flora or fauna.

Taxonomic keys, providing an example of tertiary information that is not used elsewhere.

Maps

Presence Checklists

A database of checklists is **location based**. The focus is geographic areas and the taxa that are in some way documented to be present there.

Geographical (spatial) information, providing the distribution information for the demarcated areas of interest.

Taxonomic and nomenclatural information, providing the taxon information itself.

Publishing information, providing the reference information for an observation or other documentation of a taxon's presence at a location.

Status markers and descriptors for occurrence (e.g. migrating, permanent populations, historical or ephemeral; rare, occasional, common).

Remarks and comments of the author or cited authorities.

Summary information such as species richness.

Status Checklists

A list of taxa or populations of taxa based on their conservation status or, conversely, a list of taxa or populations with their status according to one or more authority.

Taxonomic and nomenclatural information, providing the taxon information itself.

Publishing information, providing the reference information for the application of a particular status to a particular taxon or population.

Conservation status markers and descriptors (e.g. rare, endangered, or threatened).

Remarks and comments of the author or cited authorities.

Collection Notebooks

Typically **Collection Event based** or based directly on **collection number**.

Collector and collection event, noting who did the collection, when and the identifier for the collection.

Geographical (spatial) information, providing the broad location details of the point of collection.

Taxonomic and nomenclatural information, providing the collectors determination.

Environmental information, giving information of the specific locale the collection was taken.

Descriptive information may also be included

Images of the collector notebook pages

Specimen Catalogues

Specimen based, like a herbarium catalogue

Collector and collection event, noting who did the collection, when and the identifier for the collection.

Geographical (spatial) information, providing the broad location details of the point of collection.

Taxonomic and nomenclatural information, providing the collectors determination and any additional determinations since then.

Environmental information, giving information of the specific locale the collection was taken.

Descriptive information may also be included as required.

Donor information. The specimen may not have come to your institution directly following a collection event, and so details of the donating person or institution will be required.

Images of the specimens

Transaction Documentation

The focus is the **transaction event** by which objects are obtained, moved, loaned or exchanged.

Object Identifier, indicating the specific object(s) involved in the transaction.

Collection Management, noting who is involved in the transaction, their contact information and the terms of the transaction.

Taxonomic and nomenclatural information, providing the current names and any new names returned with the specimens.

Donor information, the specimen may not have come to your institution directly following a collection event, and so details of the donating person or institution may be required.

Limitations, explanation of the restrictions placed upon the specimen.

Publication information, providing the document or reference for a document that was generated based on the transacted specimen(s).

Sightings and Observations

Observation, such as bird sightings

Observer and observation event, noting who did the observation, when and the identifier for the observation.

Geographical (spatial) information, providing the broad location details of the point of collection.

Descriptive Data may also be included as required.

Taxonomic and nomenclatural data, providing the observer's determination.

Environmental data is also helpful, particularly the latitude and longitude.

Images that document the observation

Publication information, providing the document or reference for a document that formed the basis of the observation.

Project Documentation

Project based

Collector and collection event, noting who did the collection, when and the identifier for the collection.

Geographical (spatial) information, providing the broad location details of the point of collection.

Taxonomic and nomenclatural information, providing the collectors determination and any additional determinations since then.

Environmental information, giving information of the specific locale the collection was taken.

Descriptive information may also be included as required.

Process and procedural information, details of the way the project curates its specimens.

Limitations, explanation of restrictions placed upon the specimen as part of the project.

Authorization, details about permits obtained to allow collection as part of the project.

Other project related data.

Publication information, providing the document reference(s) or the document(s) that were produced as part of the project.

Though there are other possibilities not listed above, your *main* reason for undertaking a digitisation project likely falls under one of these categories. It is just as likely, however, that

what you really want is a system to handle many of these different activities. You want a catalogue of your holdings, you also want to support projects, manage transactions, keep track of your collector notebook information, etc. How are you going to make that happen?

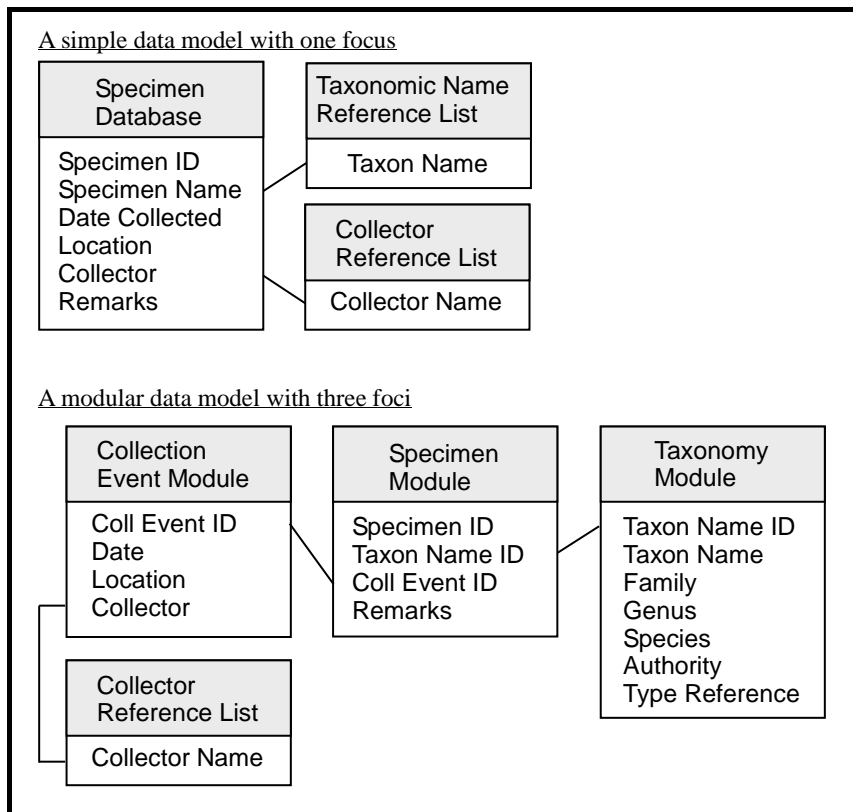
Some of these activities may be handled by simply extending a simple data model as given before for a catalogue. For example, a few fields can be added to mark whether a specimen is on loan and to whom. You can add a field for conservation status or for the project name under which the collection was made. But in all but the simplest cases, what you will need to do is develop an integrated **information management system** that allows you to organise your data and your views of the data in different ways for different purposes. You will need a data model with a modular design.

Modular Design of the Data Model:

The Information Management System

An **information management system (IMS)** allows you look at your stored information with more than one focus. For example, in a given IMS you can focus on a specific specimen record and gain access to detailed information about the collection event from which that specimen was obtained or you can alternatively focus on a collection event record and navigate to detailed information about each of the specimens collected in associate with the event. The design of an IMS is necessarily modular. In this example case, we have one module for specimens with the specimen or preparation as its base unit and a distinct separate module for collection events which has its own base unit (the collection event). The IMS allows you to focus on either the specimen or the collection event and treat the information from the other module as a type of attribute of that focus module.

In an IMS each module record consists of a base unit, in most implementations a unique identifier for each base unit, and additional information about that base unit (attributes). The data for a given attribute might be entered directly into a field in that module, or it might be selected from a reference list, or from another module. The unique identifier becomes more important in systems with modular design as can be seen in the example below for the collection event module. The base unit for a collection event is the unique combination of date, location, and collector. The only data contained in the module which reflects this base unit is the collection event ID.



An IMS with a modular data model can be a very powerful tool to help you in curation tasks and information delivery. It can also be very time consuming to develop a sophisticated data model. It may seem obvious, but with a more detailed data model there is more data to be managed and more data interactions. For example, how will you manage a nomenclature module where some of the information, the list of names, for instance, is imported from outside sources, but some of it is added in-house? How would you import an updated list of names and match it to your specimens and other in-house nomenclatural information?

Implementing a Data Model

There are two basic approaches to the design of the data model. They are the flat file approach, such as a spreadsheet or the more complicated relational database model. The advantages and disadvantages of each are considered here.

Flat file/spreadsheets

Flat file/spreadsheets have the advantage of simplicity. All you have to do is define the field names (Family, genus, collector, collector number etc.) and you can then get started with your data entry. Rapid setup time is a real advantage offered by flat file designs. Also, it is very easy to add an additional field if something has been forgotten. If data are missing you are not forced to enter any data into a particular field, which may not be the case with the relational database system. Thus data entry is usually faster using a spreadsheet, at the cost of data quality, which may cause delays in the overall process as corrections are made to the data.

This does have several disadvantages though. There is typically no form of data validation and all data is entered as text (even numbers!), which can make querying difficult. If a field

requires several values to be entered (such as a determination history), this is very difficult to achieve as the data must all be placed in one field, which makes the data difficult to search. It is also more difficult to study the data as it cannot be easily queried or filtered to present only the pertinent information.

In a spreadsheet it is possible to work around these issues to some extent and introduce some simple design constraints such as drop down lists and formatting fields to hold specific types of data. However, it is still relatively easy for the user to circumnavigate these restrictions if they so desire. It is also the case that maintenance of the drop-down lists will become an issue. In short, the more complicated the data you wish to capture, the more appropriate it is to use a relational database.

Relational Database Management Systems (RDBMS)

Relational Database Management Systems (RDBMS) can be created in a similar manner as a simple spreadsheet. You create a table in which you define your fields by name and also by data type. You can also define which fields must have data in them, without which the record will be rejected. Once this is done, you can then open the table and start entering data. This method has most of the same issues as the spreadsheet approach, but allows a slightly greater degree of control of the entered data, at a slight cost in setup time.

However this is not all that a RDBMS system can do. Lookup lists can be created and maintained via the use of additional tables which are linked together via fields known as keys. Keys are defined in two parts. Firstly there is the *primary key*, a unique value, usually a sequential number, assigned to the value you are wishing to define. This is then used in your main table as a *foreign key*, which is simply a field designed to hold the primary key of the data you are looking for. This is where the 'relational' part of the RDBMS comes in, as the primary and foreign keys create relationships between the various tables. In a similar way, the problem of multi-valued fields can be solved as a separate table with its own fields, linked together by key fields.

Using RDBMS system it is also possible to query the data in much more sophisticated ways, filtering the data to more easily find, and where necessary update, pertinent information. This makes an RDBMS approach much more powerful than the spreadsheet.

The price for this additional power is an increase in complexity of the system. It is not possible in this paper to dip more than the merest tip of a toe into the very deep pool that is RDBMS systems, so the reader is strongly advised to consult your IT staff or study database design for your system before getting started. This naturally adds considerably to the setup time required for a project but the increase in data accuracy is often worthwhile. Also, the more complicated the RDBMS becomes the more difficult it is to represent the information in a user-friendly fashion. Although Microsoft Access is capable of integrating a lookup into the users view of a table, many systems are not and designing a suitable user interface then becomes an additional requirement.

Object Orientated and Object Relational databases

Object Orientated and Object Relational databases are slowly becoming more available. It is still to say that these have similar strengths and weaknesses to the RDBMS options and should be considered in the same way as relational databases.

Data model solutions vs. programmatic tools.

It should be understood that the data model is simply a representation of the way the data are stored. It should reflect the data that you require for your project but may not reflect the way the data is actually entered or maintained. You will require some additional tools, such as a find and replace system, to keep your data up-to-date and to ensure the data is as up to date as possible. For the simple spreadsheet approaches these are already built into the system you are using. To some extent this is also true for RDBMS systems, but it is usually better to add a user-friendly front end to the system. This will generally contain forms for data entry, querying and retrieval. It will also typically include tools to calculate field information and may present join together data so it is presented in a format that is easily readable. It is important that someone in your organisation has a clear understanding of the way the system works and which tools are required for the different tasks of creating and maintaining your data. Ideally, this will be written down! This person will act as the administrator or expert for your system and will help resolve difficulties should you encounter any.

Complexity

As noted in the discussion on how complex the data model should be, it is necessary to consider how data models grow in complexity as more details and refinements develop.

Multiple value fields

It is not uncommon for a single field to represent several separate entities. Collector is an excellent example. A single field could hold the primary collector or it may hold a group of collectors such as:

- T. Wajima, S. Yoshizawa & T. Kitayama

This is not particularly user friendly for the purposes of searching, so you may wish to split these out into separate values, hence:

- T. Wajima,
- S. Yoshizawa
- T. Kitayama

If the number of values is small, then it is possible to hold each of these in a separate field in a flat file, calling the fields “collector 1”, “collector 2”, “collector 3”. However, this soon becomes unwieldy and requires many fields which may not be used for the vast majority of specimens – if one record requires a “collector 10” field, then all records have a “collector 10” field, even if they never fill the data in. It is better to have the data in a separate table, where only as many collectors as required need be stored. These are linked by primary and foreign keys (the foreign key being stored in the collectors table). Two records are illustrated in the following diagram, showing the first record having three collectors and the second record has only one, co-incidentally the same as one of the collectors from the previous specimen.

Specimen table	
Specimen	Name

SpecimenCollectors table		
Coll.	Specimen	Collector name

key	
1	Codium latum
2	Mastocarpus yendoi

key	key	
1	1	T. Wajima
2	1	S. Yoshizawa
3	1	T. Kitayama
4	2	T. Kitayama

Atomisation

Collector's names also neatly illustrate the issue of grouping names together. Taking our example of T. Kitayama, it could also be written as:

- T. Kitayama
- Kitayama, T.
- Kitayama
- Kitayama; T.
- T. Kitayaama

Or many other variations on the same name. This makes it impossible to present the data consistently as more variations of the name are entered. In order to solve this problem the name can be split into its various parts in a process known as atomisation, in this case initials and surname, giving the following:

Collector table		
Collector Id	Initials	Surname
1	T.	Kitayama

This information can then be manipulated to form calculated fields in a consistent manner as follows:

- "T." + " " + "Kitayama"
- "Kitayama" + ";" + "T."

This enables the data to be presented consistently in many different formats. This does take some extra time to enter, but improves the accuracy of the data and enforces a common format for all the projects data. Such atomisation should be applied as appropriate for your needs and your workflow. Notice however this does not solve all of the problems illustrated in the initial example, as spelling errors cannot be rectified by atomisation alone. This would have to be solved by the use of lookup lists and data checking.

Normalisation

Look back at the collectors table at the top of the page. You will notice that "T. Kitayama" is repeated twice in the table. Having to do this for multiple records increases the chance that a

spelling error, as seen in the atomisation example will occur. To reduce the chances of this happening, the data model undergoes a process known as normalisation. Put simply, this takes fields which have common data that may be repeated many times and places them into a separate table, which typically becomes a lookup table of accepted values. It is possible to normalise data repeatedly until it becomes redundant (for example, it is possible to separate out collectors initials into a separate list, but there is nothing to be gained from doing so), so it is recommended that normalisation is also applied with a healthy dose of realism. Normalisation allows data to be maintained more easily, as changing one value in one table can correct errors in many records. It is also possible to remove spelling errors by correcting the line, or by altering the records to point to the correct entry in the lookup table.

This is applied in practise once again via the use of keys. For a single valued field this is straight forward. In the following diagram, there are two specimens where the type is defined from a list in the type table. In this case both specimens are defined as isotypes.

Type table	
Type key	Name
1	Type
2	Isotype
3	Kleptotype

Specimen table		
Specimen key	Type key	Name
1	2	Codium latum
2	2	Mastocarpus yendoi

In the case of the collectors example the same principle applies. Extract out the collectors to form a table of its own, using the keys to link the two table together, providing exactly the same results as the first example.

Specimen table	
Specimen key	Name
1	Codium latum
2	Mastocarpus yendoi

SpecimenCollectors table		
Sp_Coll. key	Specimen key	Collector key
1	1	1
2	1	2
3	1	3
4	2	3

Collectors table	
Collector key	Collector Name
1	T. Wajima
2	S. Yoshizawa
3	T. Kitayama

This gives a table that is simply a list of numbers, which may seem difficult to comprehend and in truth this is the case. It is at this level of complication that the use of forms becomes necessary to hide this complication, presenting the data in a user friendly format.

It is also true that the more separated out the data is, the more difficult it is to query, as the various table must be joined together correctly to create appropriate results. The difficulties of joining tables together correctly is beyond the scope of this paper and the user is strongly recommended to study the details of their database solution before starting into this level of detailed design. Finally, the more the data is spread out over many tables, the longer it will take your computer to process the information and present it to you.

Combining multiple field values, normalisation and atomisation.

Putting the previous examples together gives three tables which look like this:

Specimen table		SpecimenCollectors table		
Specimen key	Name	Sp_Coll. key	Specimen key	Collector key
1	Codium latum	1	1	1
2	Mastocarpus yendoi	2	1	2
		3	1	3
		4	2	3

Collectors table		
Collector Id	Initials	Surname
1	T.	Wajima
2	S.	Yoshizawa
3	T.	Kitayama

As can be seen this is quite complex, however, it does illustrate the point that the more complex the system is, the more you are relying on someone who understands the details of what the system is doing and how, when it is presented to the user there is a great deal going on behind the scenes to make the database work correctly.

Section 6: Deciding on a particular database solution

In this section, we only briefly look at the various areas you should consider when selecting a database. It is highly recommended that you look at the preceding sections before selecting a database for yourself. Whatever you do, ensure your database fits your institutions IT capabilities. Picking a database you cannot actually use is probably the quickest way to guarantee your project will fail! If you do need a database with extensive IT infrastructure

requirements, you will need to include the resources to extend your IT systems, which will make your project significantly more expensive.

Whichever database design you choose, don't forget the human element. One of the most important, if not the most important, aspects of the database system you select is the user interface. If it is difficult to enter data, then data entry will inevitably be slower and your staff will be less happy in their working environment. The easier a database is to use the more widely accepted it will be and the faster data entry will proceed. This is a very difficult aspect of data entry systems to evaluate without proper testing with your real specimens and is one of the reasons why practical testing of your chosen database is considered so very important.

There are many collections databases already on the market (Berenson et al, 2003), at a wide range of scales from individual collections to full scale multi-collection institutional databases. Fortunately, databases also come at a wide range of prices and it is likely that you will find something that will fit your project funding. However, most commercial databases are also quite generalised, so you may need to consider ways of adapting the database to fit your particular needs. Should a database be close to but not exactly match your requirements, it is often possible to hire a contractor to further modify the system to reflect your needs. It may even be that you need to create a completely new database, which may not be the fastest or cheapest option, but it does give you the ability to specify exactly what you want.

When you do select your design, don't forget how complicated real world data can be. A simple example is the recording of the date a collection is made. Viewed naively, a date is a simple thing with a day, a month and a year. Collectors though seem to have a perverse wish to make things complicated and may only record the month and year, only the year, or a range of dates. Even if the date is recorded as 5/9/1815, does this mean the fifth of September 1815 or the ninth of May 1815? Reference to the collector may make this clear but that does require additional research. Dealing with data is discussed in Section 4.

The design you elect to use is referred to as the data model. For many institutions the underlying data model will depend on the commercial database package selected. If you are designing your own database, then you will be able to specify your own data model. Data models play a role in how data is disseminated outside of the institution's database, so it is highly important. Data models are discussed in Section 5. It may also play a role in migrating data into the institutional database, if this is a requirement on your project. Many data models exist, although the current emphasis is on data exchange schema such as ABCD¹ and Darwin Core². If you are designing your own system it may help to gather your own ideas together by looking at these schemas.

Dissemination of data is another issue. Some commercial database packages will have mechanisms to allow external (typically Internet) access to your data. Does this fit your requirements, or will you need to create a separate application? Separate applications do have the advantage of being designed precisely for your own requirements but equally have additional costs associated with them. Allowing access to the outside world also brings its own risks, especially if you are intending to make use of data entry online. Internet attacks by so-called hackers are a significant risk when exposing your data (Morris, 2005) and appropriate measures should be taken to guard against such attempts.

¹ See <http://www.tdwg.org/standards/id/81/>

² See <http://digir.sourceforge.net/schema/conceptual/darwin/core/2.0/darwincoreWithDiGIRv1.3.xsd>

Once you have selected a candidate database, you can then begin to consider how you can practically implement your project, which is discussed in the following section.

Which database(s) should you use?

Put simply, you can either get an existing software package from somewhere else or you can build it yourself. There are advantages to either of these approaches as well as drawbacks. Keep in mind that no solution will be perfect. There are things it will do well, things you wish it did better, and things it won't do at all. However, with appropriate planning you should be able to obtain a solution that meets your needs reasonably well and that represents a good investment of your time, money, and resources.

Existing Packages

Existing packages can either be commercial or open source. If you have the resources to pay for it, you may find that the upfront and possibly continuing expenses of a commercial package represent an appropriate expense to get you up and going rapidly. Much or all of the design work has been done for you already. The same may be true for an open source package which may have little or no cost. In either case, it is important to look for access to technical support and documentation. How active is development; what is implemented now, and what is planned for the future? How often are new versions released and what will be involved with migrating to the next version of the program? Who else is using the program and how likely is it that the program will continue to be supported in the future? But most importantly, does it do what you want it to do and expect it to do?

One of the main advantages of using an existing package is that you should be able to evaluate it prior to a full commitment to using it. Try it out if you can or at least study the demonstrations and documentation fully before committing to use it. Keep in mind that the representatives of a given package are unlikely to focus on the limitations, shortcomings, and flaws of their software. As much as possible, you need to discover these yourself during your evaluation process. Do not assume that the package will do things it is not documented to do and ensure that the functions you are interested in actually work the way you would expect them to.

It is extremely unlikely that an existing package fits exactly what you need for your situation regardless of the sales pitch. It may be too simple and may not be able to support the full range activities you expect from it. Alternatively, it may be too complex, requiring more expertise and resources to manage and maintain than you have available. It may have features you will never need, but which you will nevertheless need to maintain and interact with just to use the program. Or it may simply have the wrong focus. You could use a package that is primarily designed to database photograph collections to database your specimen collection, but is this probably isn't the best way to go.

Existing packages should always be evaluated with respect to flexibility, customisation and ad hoc solutions. At one extreme, a package may represent an abstracted system for holding a data model with built in capabilities for views, reports, searches, etc. To use this system, you will have to do a substantial amount of development and possibly programming and its use may represent little savings over designing the whole system *de novo*. On the other hand, a system may be so rigid that it simply cannot be modified to meet the particulars of your situation or may only be modified with substantial additional cost that were not foreseen upfront. This may be particularly true for commercial systems that are highly proprietary.

Between these two extremes is support for ad-hoc customisation. Many packages, for example, have placeholder fields that you can name and use for data that were not expected in the original design. Such fields, however, may not be as easy to enter into, search, control entry, export, or even view. Furthermore, these changes may not be compatible with the underlying or intended data model, making their interpretation less than obvious.

It is beyond the scope of this paper to critique existing packages or even to give a detailed listing of them. GBIF commissioned a survey of existing publicly distributed collection management and data capture software solutions (Berendsohn et al 2003) and this may be a good place to start. Other starting points include:

- **TDWG Subgroup on Biological Collection Data: Software for Biological Collection Management:** <http://www.bgbm.org/TDWG/acc/Software.htm>
- **GBIF Links to Software and Tools:** <http://www.gbif.org/links/tools>
- **Digital Taxonomy: A Web Resource for Open Source Biodiversity Informatics:** <http://digitaltaxonomy.infobio.net>
- **A searchable list of herbaria that are databasing their collections:** http://www.cals.ncsu.edu/plantbiology/ncsc/type_links.htm

There are many more programs in use than are currently listed in any of these sites, although it is possible that in the future they may be more complete. As you explore packages that may be of use for your situation, you would do well to both check with existing users of any package you are interested in and also check with other collections that may be similar to yours to see what they are using.

Building Your Own Solution

If your databasing needs are relatively simple or your resources scant, it might be worthwhile to build your own solution. A collection of just a few hundred or even a few thousand specimens can be catalogued in a flat file or spreadsheet solution that is built in less than a day. Some simple preparation, consideration of quality control issues, a look at the ABCD and/or DwC schemas (see Standards in the previous section) and you could rapidly be on your way to building a system that could be of great use to you and potential users of your data. Building your own solution could also be a good starting point. If done carefully and thoughtfully, you will likely be able to migrate your data later into an existing software package or scale up your solution with subsequent modifications should your needs or access to resources change in the future.

Alternatively, you may have the needs, resources, and access to expertise to build a more sophisticated solution. Building it yourself will give you the flexibility to tailor your solution better to your specific needs. If you have the computer and programming expertise and the time to do it, then truly building it yourself from scratch may be the best way to get just what you want. But you should be careful that it is easy to underestimate the amount of time and effort it may require to design and build an elaborate information system. Another solution would be to customise an existing package to meet your needs. If you can get hold of an open source package that is sufficiently close to what you want and which allows sufficient customisation to let you do what you want, then this may also be a satisfying route to follow.

In most cases, “building it yourself” really means getting someone or a staff to build your database or information management system for you. Depending on your situation, this can range from as simple as getting a motivated and sufficiently proficient graduate student to build it to more substantial investment, either by contracting with a commercial firm with

existing staff and resources to design and implement your solution or by hiring such staff yourself. Alternatively, your institution may have existing IT staff that you can commission or task with the development and building process. There are two important considerations to keep in mind if you follow this route. First, the quality of what you get may depend largely on your ability to communicate what it is you want and need. If you can't articulate what you need from an information management system or don't have the time to convey this to the developers, you will be hard pressed to find the suitable expertise to build what you want anyway. Secondly, it is very important that your solution developers appreciate the nature of biological data and museum collections data. Database programming expertise is simply not enough. Time and time again, I have seen solutions built that work well in theory, but then collapse when faced with the reality of the data we work with and the real world work flow their solution is supposed to enhance. When this happens, strict database developers will often try to convince you to change the data to fit the data system they have built. In other cases, they will try to "solve" the data problem with elegant database solutions that invariably are not so perfect as they had intended or, more likely, are never finished, leaving you with nothing but wasted time, effort and money.

Morris (2005) discusses many of the issues involved with relational database design and implementation in the specific context of biodiversity informatics. This is certainly a worthwhile document to consult during your planning phase and also to recommend to your development staff.

One of the advantages to a build it yourself solution is that the result should be more transparent and easily modified than if you were to use a commercial product. This is not always the case. Depending on your expertise and the manner in which it is developed, a home-grown solution can be just as much a black box as any other solution. Furthermore, the question of how it is maintained and upgraded as necessary over time needs to be considered in advance. Will it still work when a new operating system comes along? Is the source code available and accessible or just the compiled version? Finally, what if any effort will go into generating documentation? Should there be no documentation or little for your database, you should strongly consider how new workers will be trained in its use should there be turn-over in the experienced staff.

What are characteristics of a good database solution?

Whether you are looking at commercial packages, open source software, or considering building your own solution, it is important to understand the criteria with which to evaluate proposed systems. Ultimately, a "good" solution is one that both matches your needs and which works well. The following list of questions is intended to help develop your criteria as you evaluate which particular solution might work for you.

What does it really cost?

Cost includes the initial price for development, hardware cost, licenses, maintenance, upgrades, and additional software requirements. Cost also includes access to the expertise to keep the system in working order.

Is it stable?

A good database solution should be stable in three ways. First, it should work for most of your purposes now with some, but an overwhelming number of features on the "coming soon" list. Be very clear about what the proposed software really does versus what it could

do. Be very clear about how modifications and upgrades will be implemented, particularly with respect to how they impact workflow. Also evaluate how changed or added features will be tested. Will they be pretested with realistic sample data or will you have to find out if they work while you are trying to enter real data and hope for the best?

Secondly, is the software bug free? How hard is it to crash the program? Of particular interest is determining if or how likely it is to corrupt the data if the program crashes. If you do find bugs in the program down the road, how likely is it that you can get them fixed? Are the data backed up as part of the package or will you have to have an external back up system?

Finally, what is likelihood that your program will be supported over future changes in computer architecture, operating system, database program, networking protocols, and programming language? As rapidly as technology is changing, it is realistic to assume that your database solution will have a finite lifespan in its current form. Hopefully, this will be measured in years, but it is possible that some solutions will be out of date in a matter of months or even out of date at the time they are built.

Does it have good documentation and/or technical support?

Even the most powerful, elaborate information management system can be rendered useless if you can't figure out how to use it. Documentation should cover both what the solution can do and how to use it. Are tutorials included? Technical support should include issues with setting it up, how to use it, and dealing with and reporting potential bugs. Make sure to look into how technical support is accessed, how available it is, response time, and how much it costs initially and after the introduction period. Look at your needs for documentation and support for the whole package from the computer hardware to the backend database, the front end, and the implementation. You should also be able to view the data model in some form or another.

Does it have the performance you expect?

Few things are more frustrating than having to wait endlessly while the computer cranks through even the simplest function. There are many reasons for slow performance. It could be that your computer or server is slow (low processing speed), the hard drive is full, or your memory used up. Or it could be that your computer is plenty fast enough, but the program is still just plain slow. Poor programming can lead to slow performance because some routines are inherently less efficient than others. Some solutions may have background activities, such as record change tracking or validation processes, that slow down performance. Some implementations may need to periodic maintenance events, like manually rebuilding the indices or "vacuuming" the database, to keep performance from degrading. If a solution allows or requires network connections as part of its functioning, you need to also address the reliability and speed of your network while determining if a particular solution will work for you.

Some performance issues are scale dependent on the amount of data in the system. A particular solution may work great, for example, when you have only 2,000 records in it, but when you get to 200,000 records, you find that searches, imports, reports, or exports take forever.

What is the learning curve?

Few programs are going to be usable straight out of the box. You will have to have training and/or access to documentation to use it properly. In general, the more functions a program has, the longer the learning curve is to use it. But there are other factors that affect the learning curve. Some designs are more intuitive than others. How navigation and functionality rely on buttons, menu items and keystroke commands will affect the learning curve. How easy is it to search for or view data? Do you have to know SQL to perform searches or write scripts to generate reports? How similar is the functionality of a given program to programs you are already familiar with?

Long learning curves can lead to frustration even when the documentation is good and the proposed program is ultimately a reasonable solution to your needs. One of the hardest parts of evaluating a proposed solution is determining whether it is worth the time and effort necessary to learn to use it, even if it is a “good” program.

How do you initially populate the system with data?

You need to think, up front, about how all the necessary data will get into your system. Generally, our main focus will be information about specimens, but it could also be collection events, projects, publications, nomenclature or any combination of these or other types of information. Which if any of these types of information need to be entered before other types of information can be entered? For example, do you need a list of collector names or taxonomic names before you can enter a specimen? If so, where are you going to get these data? Are some data supplied with the system? Which kinds of information does the system treat as reference (i.e. static lists of values for a drop-down list) and which are modifiable and if so, by who and how? Which data can you get from existing sources and which do you need to compile yourself?

How do you enter, edit, view and delete data?

Look carefully at how these four functions occur throughout the system. Are they separated, for example, such that entering data takes place in one way, but editing existing records occurs differently or in a different view of the data? Who can delete data and how easy is it to do by accident? How are links maintained between modules when information is entered, changed, or deleted in one? Which information is required to be entered for a valid record and what happens if you have incomplete information? Are there shortcuts to help enter data that remains constant over many specimens? How easy is it to view data in unique combinations?

Can you navigate easily around the program?

A sophisticated relational database with multiple views of information and modules for various functionalities can be very powerful if you can use it. Ease of navigation is an important feature of usability. Data entry screens should allow for intuitive tab order and generally easy flow from field to field. Typically, there are more fields for entry or view for a given record than can fit on a single screen, at least with a readable size font, so look at how you navigate to see more complete information for a record. How do you move between views of a single record and lists of multiple records? How do you move from one module to another when the modules are linked or when they are distinct? For example, how do you access more information about a taxonomic name, a collector, a location or a collection event when you are looking at specimen records? How do you navigate between data entry

functions to label generation or transaction recording modules? If the database has simple, intuitive, or even sensible navigation it will improve your day-to-day satisfaction with the program. Clunky or seemingly random navigation can, at best, increase the learning curve substantially and at worse be a constant headache that reduces productivity substantially.

How does the solution improve data quality?

For almost any solution more sophisticated than a spreadsheet you will be looking for features that improve the quality of your records. This can entail drop-down choices for some fields, calculated fields that allow data to be used in different ways without re-entry, and field-level validation to ensure that entered data meet some minimal expectations for that field.

The alternative side to this is that if the program expects higher data quality than exists, can you enter lower quality data? For example, perhaps the collection date field prevents you from entering 10/32/1964, but what if the collection date is “summer 1964;” will you be able to enter this or will it force you into some ad hoc solution? Ideally, a good program walks a fine balance and will disallow or call attention to data entry errors, but also allow over-ride or alternative entry methods for lower quality data.

Another aspect of data quality involves a built-in validation process and/or support for a quality control process. Can lower quality data be identified and retrieved for post-data entry review? Can records be marked as having been subject to administrative or expert review? Can an administrator associate particular data entry issues with particular data entry personnel? Can an administrator determine when or even if a record has been changed and determine quickly which information was involved in the change?

What kind of importing functions does it have?

Programs can differ widely with respect to expectations that data will be entered directly into the database or whether information does or can come from an import from some external source. If you have substantial legacy data how will it get into the program, how will it or must it be formatted prior to import? Is import addressed only as an initial function of setting up the new solution or will the program easily support imports of new data in the future. Some programs go even further, expecting data entry to occur outside the program into a format that is then imported. Some provide stand-alone data entry modules that can be distributed to collectors. In any of these situations, it is important to review where data quality issues are addressed for imported or potentially importable data. Will you have to address them prior to import and, if so, will you have any tools to help evaluate the data quality outside your main program? Will you have to address them during import using import logs or error reports to key you into issues with the data? Once in the system will you have markers to distinguish records as belonging to a particular import set? Are there tools to improve their quality once in the system or even remove them if errors are too great?

What kind of exporting and reporting functions does it have?

Exporting and reporting functions can also vary widely among programs. Pre-packaged reports may be included to facilitate some commonly expected output, but there should be some capability to output unique reports. Look at how reports may be formatted for printing on paper if this is important to you or, alternatively, if functions that allow print-outs can be used to generate electronic output as well. Look at the support for alternative types of output

including pdf, html, xml and choice of encoding. Can you export your data in UTF-8 or ISO-8859-1 or just ASCII text? Or can you even tell what encoding you are using during export?

Support for web-based access to data is important for many database users today. At one extreme all or most interaction with the database may be through a web interface, from data entry to view access for the public. All-in-one solutions may have a substantial impact with respect to security and performance. Alternatively, the main database may interact with or export data in a format suitable for web presentation on a separate server. While this solves some problems, it is important to address the additional requirements that will be involved in developing and maintaining the web presence.

There is a lot of interest today in allowing automated or semi-automated harvesting of data such that your database can act as a node in a confederacy of like-minded databases. At the bare minimum, this entails the needed capability to export into a known schema such as ABCD or Darwin Core. Most likely it will also involve support for a data portal such as DiGiR or TAPIR and the necessary scripts to maintain a current, refreshed and compatible view of your data. Some programs may package such capabilities with their program or provide support for developing these extensions yourself.

Exporting and reporting data from a relational database can become increasingly complicated when the underlying data model is complex. This can be manifested in slow export; preparing a flat-file Darwin Core export from some programs can take as much as 24 hours of processing time even when the number of records is relatively low. It can also mean generating a report or export, on the fly, is simply a dauntingly difficult task to do. Canned reports or exports are helpful, but make sure you become aware of the specifics involved with generating a novel report or export and any tools, scripts, languages, etc. that will be needed to carry it out.

What support for networking and multiple access does a solution have?

Some solutions may reside on a single computer allowing only a single local user at a time, although this is becoming less common. More sophisticated solutions keep track of multiple users or can give different classes of users different rights. With larger projects, look for support for multiple, simultaneous access so that more than one person can be entering data at the same time or, for example, so that your collection management can be using it to process loans from her office, while data entry is occurring elsewhere. It is generally a bad idea to have multiple copies of the database in circulation such that one is the “real copy” while others are distributed for others to do searches and reports and other functions.

Can you, and if so, how do you customise it?

To one degree or another, unless you build your database completely yourself, your solution is going to be something of a “black box.” There are going to be some features that you simply do not understand how they work (but hopefully they do work!). There are going to be some things you are not going to understand why they work the way they do, but probably some trade-off is involved. If X worked the way you wanted it to, maybe Y wouldn’t work so well or at all. A big part of the evaluation period and the learning curve is finding out what you have to get used to in order to use a program properly and which things are actual deficiencies of the program.

In any case, at some point changing needs or a growing understanding of your needs is going to entail some customisation of the way a program works. It may need slight tweaking to meet the needs of your particular workflow. It may need drop-down values that are

compatible with your legacy protocols. It may need whole new modules of functionality to be added when more resources are available to add them. To the degree possible, you should determine in advance how flexible a proposed solution will be to such changes and what built-in features it has to allow customization. Can fields be reorganized on a particular view? Can you alter the tab order among fields? Can you create new views for certain tasks, and if so how? Can you modify the underlying data model or do you just change the way you interact with a relatively static and immutable model?

Custom built or open source programs may have the opposite problem, that it is too easy to inadvertently cause a change that disrupts the program function. Can a data entry person change a script underlying global navigation or delete a view?

Does it have the right focus and features for you?

A program can work beautifully, but if it isn't right for you and your needs, it is not going to be a good solution for you. Following the discussion in Section 5, a prospective program should focus on what is important to you. If a program is designed to handle any type of museum objects including artwork, architecture, and anthropological items, maybe it is not right for your collection of pinned insects. If it doesn't allow appropriate customization, you could find yourself spending most of your data entry time navigating around fields that have no relevance to the information you wish to record. Or worse, you spend all your time putting your information into a data model that doesn't allow you to access and retrieve it in the way you expect and need to in order to fulfil your mission.

Maybe you've found a program that allows you to record detailed information about your specimens and collection, but what you really need is something to keep track of all the details of your various projects and the literature you produce in conjunction with these projects. If the program doesn't focus on projects as you would like, you may find all of your efforts coming short of the outcome you expect.

As you evaluate existing solutions or prepare to build or have built your own solution, enumerate the data you are interested in entering, maintaining and having in your system. What is the focus of your information system or will it have multiple foci? For your objects of interest (see Section 4), what is your primary object information and what is secondary? What information is ancillary to your primary object information and what will be used as reference information?

There are many features that an advanced information management system may have or that you may be interested in them having. Some common features and data issues of interest to our community are discussed below. Depending on the nature, complexity, and focus of a system, these may be handled in different ways which are not necessarily better or worse than another way. It is important that you address, during evaluation or system development planning, the specifics of how these will be handled.

Handling nomenclatural information

Nomenclatural information is incredibly difficult to translate into data in a database system. The taxonomic name is, in one sense, just a label or attribute for the specimen; as in, X is a specimen of "Y" taxon. Even this is subject to interpretation, however, as there is the name as given on the label, which may or may not be spelled right and may or may not have a consistent format with respect to the authority or ranks given, and then there is the "accepted" version of the name cleaned up to match some expected or more correct format.

Parsing a taxonomic name is not necessarily straightforward. The name as given on the label may be at any of a variety of taxonomic precisions from something like “Unknown Arthropoda” to “*Rosa alba* subsp. *alba* forma *angustifolia*.” For some taxonomic groups, “species” names are more or less straightforward using binomials or trinomials, but others, particularly plants, are much more complicated. Subspecies and variety, for instance, may be at the same rank (i.e. *Rosa alba* subspecies *alba* or *R. alba* var. *alba*) or at different ranks (*R. alba* ssp. *alba* var. *ternata*). Hybrid taxa and cultivated taxa add additional complications as does the inclusion of nomenclatural authority.

There can also be information in the label name that is not strictly part of a name such as “*Rosa fulva* (sp nova?)” or “conforms favourably with. *Rosa Alba*.” A system may or may not even allow you to enter such information and if you can enter it, it is less than straightforward how to relate it to the same name without the accompanying modifier.

Then there is the issue of taxonomic hierarchy, as the label name represents some identification within a hierarchy of linked names up to Kingdom. The placement of a species within a particular hierarchy is subject to interpretation and there are generally multiple hierarchies that might be applied to the same species. These hierarchies do not necessarily follow the same rank structure either. If that is not bad enough, in most taxonomic hierarchical systems, there are always some taxa whose placement is uncertain and thus left unlinked to next higher rank.

Synonymy is a large issue. Names can be related to each other through hierarchy, but also as partial or total equivalents. Handling synonyms means not only holding more names in your system, but also maintaining the relationships among them and developing mechanisms for manipulating them and distinguishing and applying currently accepted status consistently throughout they system. A related concern involves common names. A given system may or may not allow one or more common names to be associated with each scientific name. Maintaining these represent an additional management burden.

The way a given solution handles nomenclature may be tied to expectations about how many potentially applicable names there are. A drop-down list might work well for a database of the mammals of Iowa, but it would never do for the arthropods of North America.

Type specimens add in additional complexity in that name for which the specimen is a type may or may not be the same as the label name or the currently accepted name for that taxon.

All of these considerations should make it clear that obtaining or developing a perfect solution to handling nomenclatural information is not realistic. Instead, look for a solution that handles your needs relatively well, has reasonable flexibility to handle unusual circumstances, and which does not add an unduly large burden to manage.

Tracking nomenclatural changes

The label name is subject to interpretation and later revision either by a later expert review of that specimen (“determination” or “annotation”) or by application of a new taxonomic interpretation of that name as codified in a published manuscript or treatment (“nomenclatural update”). A given name may also simply be misapplied or mistyped during data entry and, thus, need to be corrected. For many reasons, it may be useful to track changes to the names that are applied to a specimen and track the date and person who changed it. It may be useful to track why the change was made, for example, to distinguish between a typographic correction and an expert determination. It may be useful to keep track of all determinations or just the original and the most current one.

Generating labels or tags

Collection management programs often generate labels or tags for specimens, although this is most commonly found for herbaria. Because of their size, arthropod pin labels often contain abbreviations and other shortenings of the total information such that they are difficult to generate automatically from the databased information. If the collection management program allows label generation, careful consideration should be given to the interaction with workflow such that new labels get associated with the specimen that was databased and that the new labelled specimens are clearly marked as distinct from labelled specimens that still need to be databased.

Tracking curation and transactions

Curation information can be as simple as maintaining the current physical location of a specimen or it can represent a detailed track of an object through elaborate processing phases up to and including its final placement in the collection. Complications arise from maintaining proper identifiers and relating them through the processing phase. An accession may comprise readily identifiable and distinct objects at the outset, but in many cases, partitioning, sorting and later, more detailed identification may be involved before the objects are placed in the collection.

Transactions involve keeping track of loans, exchanges, and specimens sent out for identification. Tracking transactions entails information about people, institutions, policies, and documentation. In some systems and models, transactions can include transfer of a specimen from one collection to another within the host institution (such as to a teaching collection), gifts to or from the host institution or to keep record of specimens that have been deaccessioned or lost.

Marking sensitive data

Records may be need to be marked as sensitive for a variety of reasons such as for species that are rare or in danger of commercial collection or because they represent vouchers for ongoing research. Marking records as sensitive allows development and implementation of data restriction policy which, in turn, determines what to can be done with sensitive data. Data restriction policy can apply to either whole records, to certain types of information within records, or to some combination of record level and field level data access.

Marking records as sensitive can be as simple as putting in a sensitivity check field in your object record table, but more elaborate systems maintain mechanisms for recording notes multiple sensitivity markers per record and notes as to who marked a record as sensitive and why and for how long the record should be treated as sensitive.

Marking records as sensitive one record at a time can be an intensive process, particularly if the record set is large. Matching records with sensitivity criteria can also be difficult as it is hard to maintain current sensitivity information and the matching process itself can be difficult. For example, while it would be nice to mark all specimens as sensitive which are listed as federally endangered or threatened, these designations are often applied to populations, not taxa per se and most data models have a hard time capturing this detail. Cascading sensitivity from the listed name to synonyms that may appear on your specimen labels is also problematic.

Tracking record changes

An ideal system keeps track of what changed, who did it, when, and why for every field. This is generally impractical solution as it substantially inflates the data maintained by the system. A creation date/timestamp and modification date/timestamp field on the records that you are responsible for maintaining is a good start. Taxon name changes and determinations should have a separate system as the information is different. A determiner changes the name in a different way than the person who enters the determiner name.

Data exchange standards like Darwin Core expect to see a Date Last Modified field for each record. Interpretation of this field gets more complicated, however, when the exported modification date is triggered by updates to fields not in the export field set.

Allowing or supporting georeferencing

Georeferencing is the process of translating a locality description into a mappable representation of that description (Chapman and Wieczorek 2006). It is increasingly useful to be able to georeference legacy data (Beaman et al. 2004). It is also important to recognise that a georeference represents a hypothesis about where a collection event occurred. As such, for a given collection event or specimen there may be multiple georeferences. Some may represent the use of different methodologies (MaNIS vs. BioGeomancer protocols for instance) or different degrees of review (i.e., initial output vs. result of expert review or review by collector). It is also useful to distinguish between the results of a formal georeferencing and any coordinate information that came in with the specimen initially.

Handling collection dates

Collection dates tend to be problematic in information management systems because they are not always to the precision of a single day. Systems that enforce entry to a single date field should be avoided as this gives the appearance of more precision than is actually present in the data. Handling collection dates in a text field allows the information to be entered verbatim, however, the resulting data are unlikely to be useful as dates and there is likely to be a range of entered values for the same date (e.g. Aug. 23, 1976 and 10/23/1976). Dates with separators have the addition problem of ambiguity as to which number represents the day and which the month. Morris (2005) discusses the date issue at more length. Generally, a good solution will need to have a number of fields that allows entry of single dates, date ranges, and textual information (i.e. “Spring 1976”), while still allowing data to be represented as dates, at least the year information, if present.

Handling geographic administrative units

It may seem straightforward to record geographic administrative units (GAUs) with country, state/province, and county/district fields, but this is not always the case. GAUs can change such as the break up of the Soviet Republic, can change name, such as Rhodesia, and worse change geographic extent (Valencia county in New Mexico was broken into a new Cibola county and a smaller redefined Valencia county in 1978). GAUs may have different names commonly in use (i.e. “United States” vs. “U.S.A.”) and are generally different in different languages. Locality information may refer to units that are difficult to place into one of these three fields (i.e., England, the island of Hawaii, Greenland). Administrative ownership of a region may be distinct from the region itself (i.e. Martinique). Some collections do not come from any administrative unit at all (e.g. “700 miles south of Hawaii in the Pacific Ocean”)

and some come from features that define the boundary between units such as river between two states or a ridge that divides two counties.

One solution for handling GAUs is simply not to hold this information in separate fields at all, but to include it in a more general locality field. This is generally less than satisfactory, however, as GAUs are commonly used as search and retrieval criteria and allowing free entry in a locality field will entail much more keystrokes and the introduction of typographic error. Collections from more limited geographic scope may be relatively insulated from most of these problems. If your holdings are more global, it may be important to give more attention as to how to handle GUAs.

Other features and issues to evaluate

The above are only some of the features and data issues you might need to give detailed attention to as you evaluate whether a potential package or development plan is sufficient to meet your needs. Other areas that might warrant similar detailed attention include:

- Collector names and collector groups

- Separation or conflation of locality, ecological description, associated species and other collection event information

- Morphology and specimen preparation information

- Observations

- Images

- Project and voucher information

- Publications and literature

- Institutions and collection metadata

- Security and access

References

- Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* 5(1): 80-88.
- Berendsohn, W.; Güntsch, A. and Röpert, D. Survey of existing publicly distributed collection management and data capture software solutions used by the worlds natural history collections. Global Biodiversity Information Facility, 2003. http://circa.gbif.net/Public/irc/gbif/digit/library?l=/digitization_collections&vm=detailed&sb=Title
- Beaman, R., Wieczorek, J., and S. Blum. 2004. Determining space from place for natural history collections in a distributed digital library environment. *D-Lib Magazine* 10. Available at <http://www.dlib.org/dlib/may04/beaman/05beaman.html>
- Chapman, A. 2005a. Principles of Data Quality. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/DataQuality **SEE ALSO: Chapter 3 of this Manual.**
- Chapman, A. 2005b. Principles and Methods of Data Cleaning. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/DataCleaning **SEE ALSO: Chapter 4 of this Manual.**
- Chapman, A. 2005c. Uses of primary Species-Occurrence Data. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/UsesPrimaryData **SEE ALSO: Chapter 1 of this Manual.**
- Chapman, A. and O. Grafton. 2008. Guide to Best Practices for generalising sensitive primary species occurrence data. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/SensitiveData **SEE ALSO: Chapter 6 of this Manual.**
- Chapman, A.D. and J. Wieczorek (eds). 2006. Guide to Best Practices for Georeferencing. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/prog/digit/data_quality/BioGeomancerGuide **SEE ALSO: Chapter 5 of this Manual.**
- Conn, B.J. (ed.) 2000. HISPID4 – Herbarium Information Standards and Protocols for Interchange of Data, version 4 (Royal Botanic Gardens Sydney) <http://www.rbg Syd.gov.au/HISCOM>.
- Conn, B.J. 2003. Information standards in botanical databases – the limits to data interchange. *Teleopea* 10:53-60. http://www.rbg Syd.nsw.gov.au/data/assets/pdf_file/72707/Tel10Con053.pdf
- Häuser, C.L., Steiner, A., Holstein, J. & Scoble, M. J. (eds.) 2005. Digital Imaging of Biological Type Specimens. A Manual of Best Practice. Results from a study of the European Network for Biodiversity Information. Stuttgart. 304 pp.
- Lane, M. 1996. Roles of Natural History Collections. *Annals of the Missouri Botanic Garden*, vol. 83 (4): 536 – 545. <http://www.jstor.org/view/00266493/di995851/99p0266q/0>
- Losee, R.M. 1997. A Discipline Independent Definition of Information. *Journal of the American Society for Information Science*. 48 (3): 254-269. <http://www3.interscience.wiley.com/cgi-bin/fulltext/39670/PDFSTART?CRETRY=1&SRETRY=0>
- Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond integrity analysis. Proceedings of the Conference on Information Quality (IQ2000): 200 - 209. Boston: Massachusetts Institute of Technology. <http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf> [Accessed 18 February 2008].
- McLaren, S.M. et al. 1996. Documentation standards for automatic data processing in mammalogy. Version 2.0. American Society of Mammalogists. 68 pages. Available at: <http://www.mammalsociety.org/committees/comminformatics/docstandards.pdf>
- Meier, R. & R. Dikow. 2004. Significance of specimen databases from taxonomic revisions for estimating and mapping the global species diversity of invertebrates and repatriating reliable specimen data. *Conservation Biology* 18(2): 478–488. <http://www.blackwell-synergy.com/doi/abs/10.1111/j.1523-1739.2004.00233.x>
- Morris, P.J. 2005. Relational database design and implementation for biodiversity informatics. *PhyloInformatics* 7: 1-66. <http://systbio.org/files/phyloinformatics/7.pdf>
- Peterson, A.T. and Navarro-Sigüenza, A.G. 2002. Computerising bird collections and sharing data openly: Why bother? *Bonner Zoologische Beiträge* 51: 205-212.
- Redman, T. C. 1996. Data quality for the information age. Artech House Inc.
- McLeod, S. and M. C. Winans 1991. Logistics and planning for computerization. Section 2 in: Stanley D. Blum (ed.), Society of Vertebrate Palaeontology, USA: Guidelines and Standards for Fossil Vertebrate databases. 129pp.
- Snow, N. 2005. Professional Biologist: Successfully curating smaller herbaria and natural history collections in academic settings. *BioScience* 55: 771-779.
- Wheeler, Q. 2004, What if GBIF? *BioScience* 54:717 http://goliath.ecnext.com/coms2/summary_0199-48075_ITM

Appendix A: Business Case Considerations

Why digitise your collection?

- Wider dissemination of Data
- Enable your data to be studied in different ways
- Enhance curatorial activities.
- Protect your specimens.
- Aid research by reducing future transcription time.
- Fits the institutions corporate goals.
- Enhances the ability of the institution to contribute in areas beyond its traditional remit.

Identify your goals

Institutional or individual?

- Institutional* A database that must cope with a wide range of specimens and many people entering data.
- Individual* The database will only be required to handle your specimens and data standards.

Who are the principal clients of your solution?

- Individuals on a specific project
- Researchers generally
- Curatorial staff at the institution.
- Others?

What language(s) will you support?

More languages cause greater complexity.

How much data?

The number of records affects the time digitisation will take and the scale of the database required for storing the information.

What data quality?

Will you record:

- Collection data
- Taxonomic information
- Storage location
- Habitat information
- Initial description
- The specimen itself

Data capture or data interpretation?

- Data Capture* Record the data presented on the specimen as written.
- Data Interpretation* Alter the data to correct errors, such as incorrectly naming the specimen.

Enhancing existing practices in the institution

Document the ways in which digitisation will aid the curators in your institution.

Imaging

- What will be imaged?
- How will you take your images?
- How detailed will your images be?
- What format will the images be stored as (JPG, TIFF etc.).
- Where will they be stored?
- How will they be accessed?

Understand what digitisation will not do

- Databasing is not a money saving option.
- Digitising your collection will not create new information for you
- Specimens will still need to be physically stored and handled

When do you want the dataset be to be available?

- Short term.** Work that can be completed over a six to twelve month period.
- Intermediate.** Data entry over approximately an 18 month period.
- Long term.** Any project lasting longer than 18 months.

Future requirements

- How will the database continue after the current project ends?

Staffing

Who will do the digitisation?

- Curatorial staff as part of their regular work.
- External contract staff/company
- Volunteer staff?
- Visiting researchers?
- Project staff?

How many can database at once?

- One person working at a single database
- Several people using individual databases.
- Several people sharing the same database

Is expert help available?

- *Yes* Your project will run much more smoothly.
- *No* Consider how suitable expertise will be made available to your project.

Is suitable expertise available?

- Data owners
- Data experts
- Technical staff
- Project management

Limitations

Is access to your data restricted in any way

- *Yes* Note which fields/specimens will not be released and why.
- *No* Consider the possible consequences of not restricting the data.

Does your institution require you to use an existing system?

- *Yes* Record how will this be integrated into your project and if this limits you in any way.
- *No* A database will have to be selected along with appropriate data standards.

Do you have legacy data (electronic or paper)?

- *Yes* How will this be integrated into your project and will you be able to check the data quality?
- *No* You will be able to set the data quality standards and provide reasonable quality assurance.

Do you already have project deadlines?

- *Yes* Prioritise discovery of how long you will actually take to digitise your specimens, then work backwards to discover how much time you have to plan the project. If there is insufficient time, consider requesting a project extension.
- *No* Take your time to properly plan your project.

Will you be working outside your institution?

- *Yes* Document the effects this will have on your database and factor suitable travel expenses into your resource requirements.
- *No* More freedom is allowed in choosing your database.

Physical Requirements

Where will the digitisation take place?

- Digitise in the collection itself
- Establish a dedicated area for digitisation
- Digitise in an entirely different location.

Document your existing I.T. infrastructure.

- This will allow you greater security when selecting an appropriate database.

Conclusions

Is your project feasible?

- *Yes* Begin to consider ways to implement your plan.
- *No* Revise your plans until they are practical.

Do your goals exceed your limitations?

- *Yes* Consider the following options when writing the action plan:
 - Can changing working practices free up time to work on your project?
 - Can other nearby institutions help out?
 - Who might fund your project?
 - Should your project be broken down into several stages
- *No* Start writing your action plan.

Appendix B: Action Plan Issues

Which Database?

Pick a database solution

- Commercial
- Open source
- Modified commercial or open source
- Bespoke

How long will it take to build or implement? (include in the project lead time).

Resources

- How many staff do you need? (Digitisers, manager and other staff).
- How will you train your workers?
- What are you going to have to buy?
- What budget do you require?

What will your workflow be?

- Collecting and returning the specimens.
- Digitisation location.
- Data Quality.
- Adding value to the original data
- Imaging.
- Data order
- Data checking.
- Can procedures be overlapped? .
- What effect will staff absence have on your workflow?
- Are there any bottlenecks in your plan?

The human element

- Staff loss or extended staff absence
- Training

Contingency planning/risk analysis

- How will you address the issue of not making the required digitisation rate, if it becomes an issue?
- What happens if your computer breaks down?
- Backup strategies.
- Malicious alteration of data.
- What will you do to document your solution/implementation?
- Are there other risks you should take account of?

Conclusions

Will your solution provide the appropriate level of data quality?

- *Yes* Your data is perfect for your current project and should be useful in other projects.
- *No* Can you:
 - add resources to improve data, or
 - reduce the total number of specimens worked on, allowing more time to enhance the remaining data, or
 - Improve the data quality in a future project?

Does your chosen solution match your goals, limitations, and resources?

- *Yes* Implementing the project should be straightforward.
- *No* Refine your solution.

Will your solution handle your future requirements?

- *Yes* Maintaining and extending your data will be easier.
- *No* Not an issue for the current project but may be a problem for future projects.

Is your solution a good return on your investment?

- *Yes* Begin to consider ways to implement your plan.
- *No* Revise your plans until they are practical.

How long will it take to database your collection?

Chapter 3

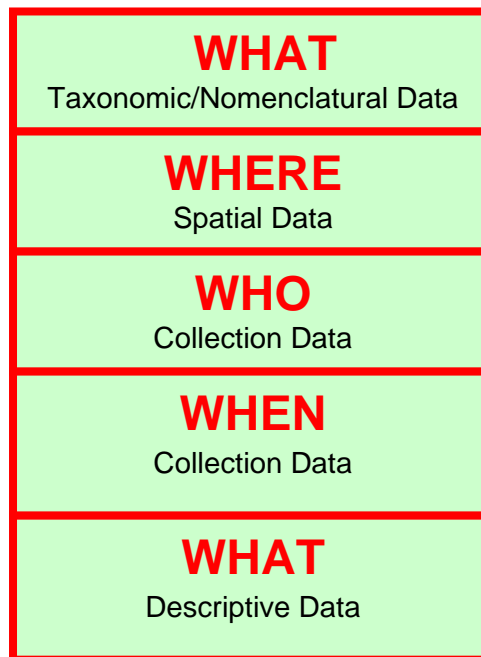
Data Quality

Introduction	1
Definitions	3
Principles of data quality	9
Taxonomic and nomenclatural data.....	22
Spatial data	26
Collector and collection data	28
Descriptive data	29
Capturing data.....	30
Data entry and acquisition	32
Documenting data.....	34
Storage of data	39
Manipulation of spatial data	43
Representation and presentation	44
Conclusion	50
References	51
Index to Chapter 3	56

This Chapter is equivalent to:

Chapman, A. 2005. *Principles of Data Quality*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 58 pp. ISBN: 87-92020-03-8 (available as a standalone PDF from <http://www.gbif.org>)

Introduction



Data quality principles have become a core business practice in fields such as business (SEC 2002), medicine (Gad and Taulbee 1996), GIS (Zhang and Goodchild 2002) remote sensing (Lunetta and Lyon 2004) and many others over recent times, but are only now becoming universally accepted by the museum and taxonomic community. The rapid increase in the exchange and availability of taxonomic and species-occurrence data has now made the consideration of such principles an important agenda item as users of the data begin to require more and more detail on the quality of this information. Indeed, some outside the museum community see the quality of museum data as being generally unacceptable for use in making environmental conservation decisions, but is this really a result of the quality of the data or of their documentation? But these data are of critical importance. Because of their collection over time, they provide irreplaceable baseline data about biological diversity during a time when humans have had tremendous impact on such diversity (Chapman and Busby 1994). They are an essential resource in any effort to conserve the environment, as they provide the only fully documented record of the occurrence of species in areas that may have undergone habitat change due to clearing for agriculture, urbanisation, climate change, or been modified in some other way (Chapman 1999).

These are some of the ideas I have tried to expand on below, as well as putting forward a number of principles of data quality that should become core to the business of museums and herbaria as they release their data to the broader community.

Data quality and error in data are often neglected issues with environmental databases, modelling systems, GIS, decision support systems, etc. Too often, data are used uncritically without consideration of the error contained within, and this can lead to erroneous results, misleading information, unwise environmental decisions and increased costs.

Plant and animal specimen data held in museums and herbaria provide a vast information resource, providing not only present day information on the locations of these entities, but also historic information going back several hundred years (Chapman and Busby 1994).

There are many data quality principles that apply when dealing with species data and

especially with the spatial aspects of those data. These principles are involved at all stages of the data management process. A loss of data quality at any one of these stages reduces the applicability and uses to which the data can be adequately put. These include:

- Data capture and recording at the time of gathering,
- Data manipulation prior to digitisation (label preparation, copying of data to a ledger, etc.),
- Identification of the collection (specimen, observation) and its recording,
- Digitisation of the data,
- Documentation of the data (capturing and recording the metadata),
- Data storage and archiving,
- Data presentation and dissemination (paper and electronic publications, web-enabled databases, etc.),
- Using the data (analysis and manipulation).

All these have an input into the final quality or “fitness for use” of the data and all apply to all aspects of the data – the taxonomic or nomenclatural portion of the data – the “what”, the spatial portion – the “where” and other data such as the “who” and the “when” (Berendsohn 1997).

Before a detailed discussion on data quality and its application to species-occurrence data can take place, there are a number of concepts that need to be defined and described. These include the term data quality itself, the terms accuracy and precision that are often misapplied, and what we mean by primary species data and species-occurrence data.



Don't underestimate the simple elegance of quality improvement. Other than teamwork, training, and discipline, it requires no special skills. Anyone who wants to can be an effective contributor.

(Redman 2001).

Definitions

Species-occurrence data

Species-occurrence data is used here to include specimen label data attached to specimens or lots housed in museums and herbaria, observational data and environmental survey data. In general, the data are what we term “point-based”, although line (transect data from environmental surveys, collections along a river), polygon (observations from within a defined area such as a national park) and grid data (observations or survey records from a regular grid) are also included. In general we are talking about georeferenced data – i.e. records with geographic references that tie them to a particular place in space – whether with a georeferenced coordinate (e.g. latitude and longitude, UTM) or not (textual description of a locality, altitude, depth) – and time (date, time of day). In general the data are also tied to a taxonomic name, but unidentified collections may also be included. The term has occasionally been used interchangeably with the term “primary species data”.

Primary species data

“Primary species data” is used to describe raw collection data and data without any spatial attributes. It includes taxonomic and nomenclatural data without spatial attributes, such as names, taxa and taxonomic concepts without associated geographic references.

Accuracy and Precision

Accuracy and *precision* are regularly confused and the differences are not generally understood. The differences are best explained through example (figure 1).

Accuracy refers to the closeness of measured values, observations or estimates to the real or true value (or to a value that is accepted as being true – for example, the coordinates of a survey control point) as shown in figure 1.

Precision (or *Resolution*) can be divided into two main types. *Statistical precision* is the closeness with which repeated observations conform to themselves. They have nothing to do with their relationship to the true value, and may have high precision, but low accuracy as shown in figure 1a. *Numerical precision* is the number of significant digits that an observation is recorded in and has become far more obvious with the advent of computers. For example a database may output a decimal latitude/longitude record to 10 decimal places – i.e. ca .01 mm when in reality the record has a resolution no greater than 10-100 m (3-4 decimal places). This often leads to a false impression of both the resolution and the accuracy.

These terms – accuracy and precision – can also be applied to non-spatial data as well as to spatial data. For example, a collection may have an identification to subspecies level (i.e. have high precision), but be the wrong taxon (i.e. have low accuracy), or be identified only to Family level (high accuracy, but low precision).

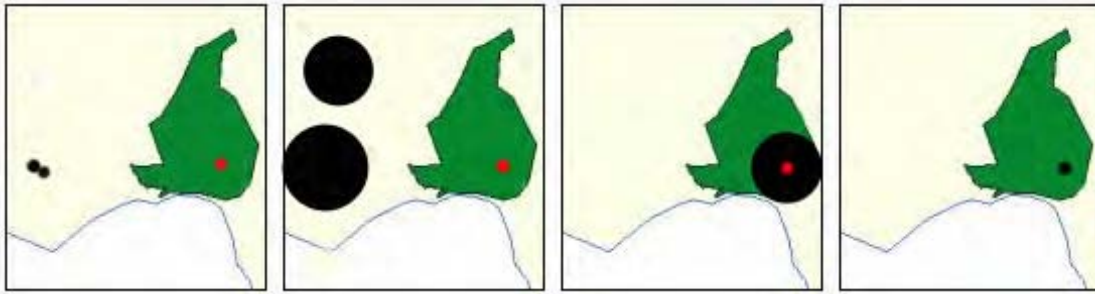


Fig. 1. Shows the differences between accuracy and precision in a spatial context. The red spots shows the true location, the black spots, represent the locations as reported by a collector.

- a. High precision, low accuracy.
- b. Low precision, low accuracy showing random error.
- c. Low precision, high accuracy.
- d. High precision and high accuracy.

Quality

Quality as applied to data, has various definitions but in the geographic world one definition is now largely accepted – that of “fitness for use” (Chrisman 1983) or “potential use”. This is the definition that has been adopted by most modern spatial data transfer standards (ANZLIC 1996a, USGS 2004). It is also being increasingly used in non-spatial areas such as in economics and business. Some (English 1999, for example) believe that the definition “fitness for use” is a little restrictive and argue for a definition that also includes fitness for future or potential uses.

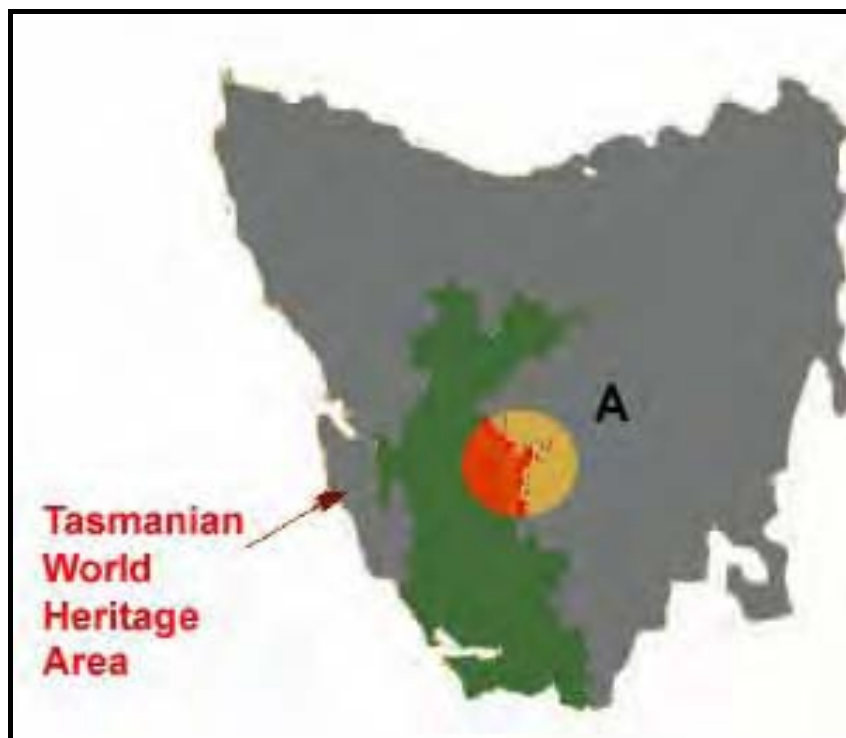


Fig. 2. Map of Tasmania, Australia, showing a record (A) collected with an accuracy of 0.5° (ca. 50 km) as shown by circle. The footprint area of possible collection (determined using the accuracy value) overlaps the Tasmanian World Heritage Area.

An example of the use of the concept of “*Fitness for Use*” can be seen in figure 2. The collection of a particular species (marked ‘A’) has an accuracy of 0.5° of Latitude (ca. 50 km). If one is preparing a list of Tasmanian species, and want to know if that species occurs in Tasmania, then the record is suitable to answer that question – the collection is “fit for use” and can therefore be regarded as of high quality for that purpose. On the other hand, if one wants to know if the species occurs in the Tasmanian World Heritage area or not, then one cannot answer that question from the record – it may, or it may not. The data are not “fit for that use” and are thus of low quality for that purpose. The latitude and longitude values in the database may be very precise and give the appearance of having a high accuracy and this can be misleading to the user of the record does not also include a value for the accuracy.

Similar cases occur with the non-spatial components of the data where a mis-identification, for example, can make the data of little value and thus not “fit for purpose”. If one is studying the distribution of a species (or its physiology or ecology, etc.), having the wrong name attached to the specimen or observation can lead to misleading and wrong results.

Data quality is multidimensional, and involves data management, modelling and analysis, quality control and assurance, storage and presentation. As independently stated by Chrisman (1991) and Strong *et al.* (1997), data quality is related to use and cannot be assessed independently of the user. In a database, the data have no actual quality or value (Dalcin 2004); they only have *potential* value that is *realized* only when someone uses the data to do something useful. Information quality relates to its ability to satisfy its customers and to meet customers’ needs (English 1999).

Redman (2001), suggested that for data to be fit for use they must be accessible, accurate, timely, complete, consistent with other sources, relevant, comprehensive, provide a proper level of detail, be easy to read and easy to interpret.

One issue that a data custodian may need to consider is what may need to be done with the database to increase its useability to a wider audience (i.e. increase its potential use or relevance) and thus make it fit for a wider range of purposes. There will be a trade off in this between the increased useability and the amount of effort required to add extra functionality and useability. This may require such things as atomising data fields, adding geo-referencing information, etc.



Data are of high quality if they are fit for their intended use in operations, decision-making, and planning (Juran 1964).

Quality Assurance/ Quality Control

The difference between quality control and quality assurance is not always clear. Taulbee (1996) makes the distinction between Quality Control and Quality Assurance and stresses that one cannot exist without the other if quality goals are to be met. She defines

- *Quality Control* as a judgment of quality based on internal standards, processes and procedures established to control and monitor quality; and
- *Quality Assurance* as a judgment of quality based on standards external to the process and is the reviewing of the activities and quality control processes to insure that the final products meet predetermined standards of quality.

In a more business-oriented approach, Redman (2001) defines *Quality Assurance* as

“those activities that are designed to produce defect-free information products to meet the most important needs of the most important customers, at the lowest possible cost”.

How these terms are to be applied in practice is not clear, and in most cases the terms seem to be largely used synonymously to describe the overall practice of data quality management.

Uncertainty

Uncertainty may be thought of as a “*measure of the incompleteness of one’s knowledge or information about an unknown quantity whose true value could be established if a perfect measuring device were available*” (Cullen and Frey 1999). Uncertainty is a property of the observer’s understanding of the data, and is more about the observer than the data *per se*. There is always uncertainty in data; the difficulty is in recording, understanding and visualising that uncertainty so that others can also understand it. *Uncertainty* is a key term in understanding risk and risk assessment.

Error

Error encompasses both the imprecision of data and their inaccuracies. There are many factors that contribute to error.

The usual view of errors and uncertainties is that they are bad. This is not necessarily so, however, because it can be useful to know how errors and uncertainties occur, how they can be managed and possibly reduced... A good understanding of errors and error propagation leads to active quality control” (Burrough and McDonnell 1998).

Error is generally seen as being either random or systematic. *Random error* tends to refer to deviation from the true state in a random manner. *Systematic error* or bias arises from a uniform shift in values and is sometimes described as having ‘relative accuracy’ in the cartographic world (Chrisman 1991). In determining ‘fitness for use’ systematic error may be acceptable for some applications, and unfit for others. An example may be the use of a different geodetic datum¹ – where, if used throughout the analysis, may not cause any major problems. Problems will arise though where an analysis uses data from different sources and with different biases – for example data sources that use different geodetic datums, or where identifications may have been carried out using an earlier version of a nomenclatural code.

“Because error is inescapable, it should be recognised as a fundamental dimension of data” (Chrisman 1991). Only when error is included in a representation of the data is it possible to answer questions about limitations in the data, and even limitations in current knowledge. Known errors in the three dimensions of space, attribute and time need to be measured, calculated, recorded and documented.

Validation and Cleaning

Validation is a process used to determine if data are inaccurate, incomplete, or unreasonable. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking of suspect records. Validation checks may also involve checking for compliance against

¹ Different geographic datums can lead to systematic shifts in the actual position (of a lat/long coordinate) of up to about 400 meters in some parts of the earth,

applicable standards, rules, and conventions. A key stage in data validation and cleaning is to identify the root causes of the errors detected and to focus on preventing those errors from re-occurring (Redman 2001).

Data cleaning refers to the process of “fixing” errors in the data that have been identified during the validation process. The term is synonymous with “data cleansing”, although some use data cleansing to encompass both data validation and data cleaning. It is important in the data cleaning process that data is not inadvertently lost, and changes to existing information be carried out very carefully. It is often better to retain both the old (original data) and the new (corrected data) side by side in the database so that if mistakes are made in the cleaning process, the original information can be recovered.

A number of tools and guidelines have been produced in recent years to assist with the process of data validation and data cleaning of species data. These will be covered in the associated document on *Principles and Methods of Data Cleaning*. The process of manual cleaning of data is a laborious and time consuming one, and is in itself prone to errors (Maletic and Marcus 2000).

The general framework for data cleaning (after Maletic and Marcus 2000) is:

- Define and determine error types
- Search and identify error instances
- Correct the errors
- Document error instances and error types
- Modify data entry procedures to reduce future errors

Truth in Labelling

Truth in Labelling is usually understood as being the documentation of quality of goods and products for sale or made available to third parties. For species-occurrence data, this will usually be comprised of the metadata, as long as the metadata fully document aspects of quality, quality control procedures and methods, and/or measured quality statistics relevant to the data. Truth in labelling is a primary function leading to certification and accreditation where these are appropriate. Most museums and herbaria already carry out this with respect to information on the expert and the date that the identification was performed (determinavit information), but this is seldom extended to other information in the record or with observational and un-vouchered survey data.

Users

Who are the users? Users of the data involve everyone at all stages of the information chain (figure 3). In the case of primary species data, they include in-house users such as taxonomists, managers, researchers, technicians, collectors, as well as the external and downstream users such as policy and decision makers, scientists, agriculturalists, foresters and horticulturalists, environmental managers, NGOs (environmental and production), medical professionals, pharmacologists, industry professionals, botanic garden and zoo keepers, the general public (including home gardeners) and community users. Species-occurrence data have endless users and involve virtually the whole community in one way or another.

Primary species data have often been collected without the broader user community in mind. Traditionally, the data, especially museum and herbarium data have been collected with the main aim of providing information for taxonomic or biogeographic research. This has been an essential process, but in today’s world the providers of funding for these institutions, often government agencies, are looking for a greater return on their dollar, and thus for the data to

have increased value through their availability for additional uses. In particular, governments are looking to use the data for improved environmental decision-making, environmental management and conservation planning (Chapman and Busby 1994), and curators of these data cannot afford to ignore these users or their needs. With good feedback mechanisms in place, users can provide feedback on data quality, and thus can be an important link in the data quality chain as discussed below.



Determining user needs is difficult and hard work. But there is no substitute for doing so and the rewards of doing so are great.

Principles of Data Quality

Experience has shown that treating data as a long-term asset and managing it within a coordinated framework produces considerable savings and ongoing value. (NLWRA 2003).

Principles of data quality need to be applied at all stages of the data management process (capture, digitisation, storage, analysis, presentation and use). There are two keys to the improvement of data quality – they are prevention and correction. Error prevention is closely related to both the collection of the data and the entry of the data into a database. Although considerable effort can and should be given to the prevention of error, the fact remains that errors in large data sets will continue to exist (Maletic and Marcus 2000) and data validation and correction cannot be ignored.

Error prevention is considered to be far superior to error detection, since detection is often costly and can never guarantee to be 100% successful (Dalcin 2004). Error detection, however, has a particularly important role to play when dealing with legacy collections (Chapman and Busby 1994, English 1999, Dalcin 2004) such as is the case with much of the primary species data and species-occurrence data considered here.



Begin by setting a data vision, developing a data policy and implementing a data strategy - not by carrying out unplanned, uncoordinated and non-systematic “data cleaning” activities.

The Vision

It is important for organisations to have a vision with respect to having good quality data. This applies especially to organisations that plan to make their data available to others. A good data quality vision will usually enhance the organisation’s overall vision (Redman 2001) and improve the operational procedures of the organisation. In developing a vision, managers should focus on achieving an integrated management framework in which leadership, people, computer hardware, software applications, quality control and data are brought together with appropriate tools, guidelines and standards to maintain the data and turn them into quality information products (NLWRA 2003).

A data quality vision:

- forces an organisation to think about its long-term data and information needs and their relation to the organisation’s long-term success,
- motivates actions in the right direction – i.e. towards quality,
- provides a sound basis for decision-making both within and without the organisation,
- formalises the recognition of data and information as being core assets of the organisation,
- maximises use of the organisation’s data and information, avoids duplication, facilitates partnerships, and improves equity of access, and
- maximises integration and interoperability.

The Policy

As well as a vision, an organisation needs a policy to implement that vision. The development of a sound data quality policy is likely to:

- force the organisation to think more broadly about quality and to re-examine their day-to-day practices,
- formalise the processes of data management,
- assist the organisation in being more clear about its objectives with respect to
 - reducing costs,
 - improving data quality,
 - improving customer service and relations, and
 - improving the decision-making process,
- provide users with confidence and stability when accessing and using data arising from the organisation,
- improve relations and communication with the organisation's clients (both data providers and data users),
- improve the standing of the organisation in the wider community, and
- improve the chances of better funding as best-practice targets are approached.

The Strategy

Because of the vast amounts of data held by large institutions, there is a need to develop a strategy for capturing and checking of the data (also see under *Prioritising*, below). A good strategy to follow (for both data entry and quality control) is to set short, intermediate and long-term goals. For example (after Chapman and Busby 1994):

- **Short term.** Data that can be assembled and checked over a 6-12-month period (usually includes data that are already in a database and new data that require less quality checking).
- **Intermediate.** Data that can be entered into a database over about an 18-month period with only a small investment of resources and data that can be checked for quality using simple, in-house methods.
- **Long term.** Data that can be entered and/or checked over a longer time frame using collaborative arrangements, more sophisticated checking methods, etc. May involve working through the collection systematically by selecting:
 - Taxonomic groups that have been recently revised or are in the process of taxonomic study within the institution.
 - Important collections (types, special reference collections, etc.)
 - Key groups (important families, taxa of national significance, listed threatened taxa, ecologically/environmentally important taxa).
 - Taxa from key geographic regions (e.g. from developing countries with the aim of sharing of data with countries of origin, geographic areas of importance to the institution).
 - Taxa that form part of collaborative arrangements with other institutions (e.g. an agreement to database the same taxa across a range of institutions).
 - Moving systematically through the collection from start to finish.
 - Recent acquisitions in preference to backlogged collections.

Some of the principles of good data management that should be included in a strategy include (after NLWRA 2003):

- Not reinventing information management wheels
- Looking for efficiencies in data collection and quality control procedures
- Sharing of data, information and tools wherever possible
- Using existing standards or developing new, robust standards in conjunction with others
- Fostering the development of networks and partnerships

- Presenting a sound business case for data collection and management
- Reducing duplication in data collection and data quality control
- Looking beyond immediate use and examining the requirements of users
- Ensuring that good documentation and metadata procedures are implemented.

Prevention is better than cure

The cost to input a collection into a database can be substantial (Armstrong 1992) but is only a fraction of the cost of checking and correcting the data at a later date. It is better to prevent errors than to cure them later (Redman 2001) and it is by far the cheaper option. Making corrections retrospectively can also mean that the incorrect data may have already been used in a number of analyses before being corrected, causing downstream costs of decisions made on poor data, or of re-conducting the analyses.

Prevention of errors does nothing for errors already in the database, however, data validation and cleaning remains an important part of the data quality process. The cleanup process is important in identifying the causes of the errors that have already been incorporated into the database and should then lead to procedures that ensure those errors aren't repeated. Cleanup must not occur in isolation though; otherwise the problems will never disappear. The two operations, data cleaning and error prevention, must run concurrently. To decide to clean the data first and worry about prevention later, usually means that error prevention never gets satisfactorily carried out and in the meantime more and more errors are added to the database.

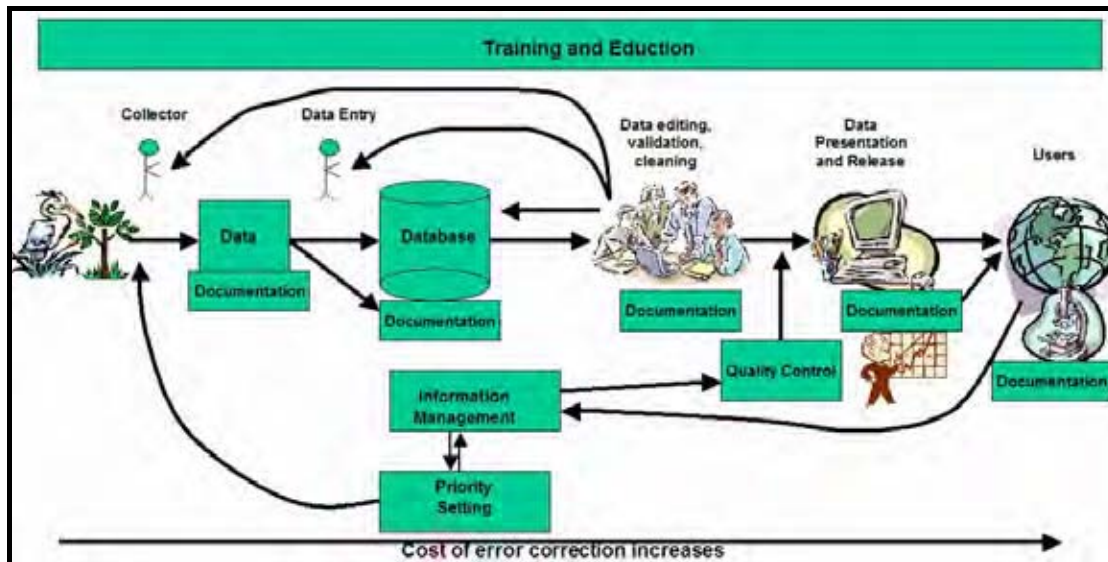


Fig. 3. Information Management Chain showing that the cost of error correction increases as one progresses along the chain. Good documentation, education and training are integral to all steps.

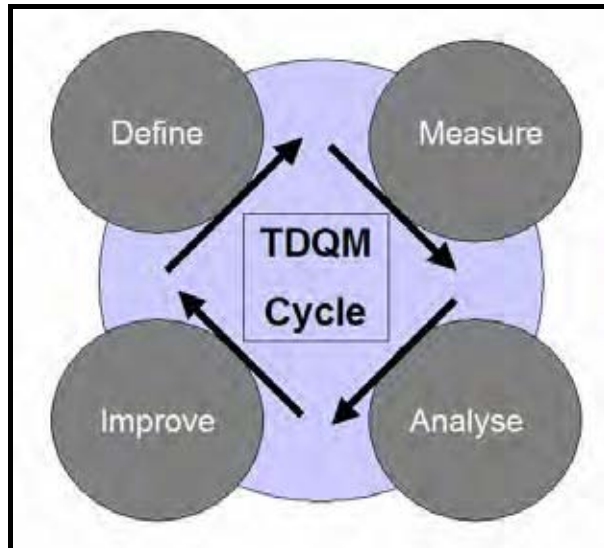


Fig. 4. *The Total Data Quality Management cycle showing the cyclical nature of the data management process (after Wang 1998).*

Custodians and owners of data (individual collection agencies such as museums and herbaria) are largely responsible for the quality of their data. None-the-less, those supplying the data and those using the data, also have responsibilities.



Assign responsibility for the quality of data to those who create them. If this is not possible, assign responsibility as close to data creation as possible
(Redman 2001)

The collector has primary responsibility

The primary responsibility for the management of data quality rests with the collector of the data. It is their responsibility to make sure that:

- label information is correct,
- label information is accurately recorded and documented,
- locality information is as accurate as possible, and both accuracy and precision are documented,
- collection methodologies are fully documented,
- label or field notes are clear and unambiguous, and
- label information is legible and readable by the data input operators.

If the information on the label or in the collector's notebook is not clear and accurate, then it is extremely difficult to correct it retrospectively. This is less important with respect to the taxonomic portion of the data in cases where voucher collections are retained, as it can, and usually is, checked by experts at a later date.

It is also important that notes on location and subsidiary information be made at the time of collection or observation and not left to the end of the day or until one returns to the laboratory as has often been the case in the past.



Most data comes into an organisation from “suppliers”, and it is much easier to develop good data collection practices than to correct errors downstream.

The custodian or curator has the core or long-term responsibility

The custodian (or steward) of the data (museum, herbarium, university, conservation agency, NGO, or private individual) has the long-term responsibility for maintaining and improving the data quality for as long as they retain responsibility for the data (see, for example, a list of responsibilities of custodianship in Olivieri *et al.* 1995, p. 623). It is important that the custodian organisation assign over-riding responsibility for managing the data quality within the organisation, but it is also essential that the organisation have a data quality culture such that every individual within the organisation knows they have a part in the responsibility for the quality of data held by the organisation. It is the responsibility of the custodian to ensure that:

- the data are transcribed into the database correctly and accurately from the collector's notes,
- quality control procedures are implemented and exercised during data capture,
- data and the data quality are adequately and accurately documented,
- validation checks are routinely carried out on the data,
- validation checks carried out are fully documented,
- the data are stored and archived in a suitable manner (see notes on storage below),
- earlier versions are systematically stored to allow comparisons and return to “uncleaned” data,
- data integrity is maintained,
- the data are made available in a timely and accurate manner with documentation that allows users to determine “fitness for use”,
- custodian responsibilities as to privacy, intellectual property rights, copyright, and sensitivities of traditional/indigenous owners are maintained,
- conditions of use of the data are maintained and made available along with any restrictions on use and known areas of unsuitability of the data,
- all legal requirements with respect to the data are honoured and complied with,
- feedback from users on the data quality is dealt with in a timely manner,
- data quality is maintained to the highest level at all times,
- all known errors are fully documented and made known to users.



Data ownership and custodianship not only confers rights to manage and control access to data, it confers responsibilities for its management, quality control and maintenance. Custodians also have a moral responsibility to superintend the data for use by future generations

User responsibility

Users of the data also have a responsibility to data quality. Users need to feed back to custodians information on any errors or omissions they may come across, errors in documentation of the data, and additional information they may need recorded in the future, etc. It is often the user, when looking at the data in the context of other data, who can identify errors and outliers in the data that would otherwise go un-noticed. A single museum may have only a subset of the total available data (from one State or region for example), and it is only when the data are combined with data from other sources that errors may become obvious.

Depending upon the purposes of data collection in an institution, the user may also have valuable contributions to make toward assisting in the setting of future priorities with respect to data collection and validation (Olivieri *et al.* 1995).

The user also has a responsibility for determining the fitness of the data for their use, and to not use the data in inappropriate ways.



Users and collectors have important roles to play in assisting custodians in maintaining the quality of the data in the collections, and both have a vested interest in the data being of the highest possible quality.

Building of partnerships

The building of partnerships for the maintenance of data quality can be a rewarding and cost-saving measure. This is particularly so with museums and herbaria, where duplicate records are often distributed between a number of museums. Many library communities use a form of collaboration and partnership to improve cataloguing of library materials (Library of Congress 2004) and museums and herbaria could easily operate in a similar manner. Such partnerships and collaborative arrangements could be developed with:

- important data collectors (in order to improve the flow of information – for example by developing standard data collection and reporting forms, provision of GPSs, etc.),
- other institutions holding similar data (e.g. duplicate collections),
- other like-institutions with similar data quality needs and that may be developing data quality control methods, tools, standards and procedures,
- key data brokers (such as GBIF) who provide a role in collating and distributing information from a number of data providers,
- users of the data (especially those that may carry out validation tests on the data during or prior to analysis), and
- statisticians and data auditors who may be able to improve methodologies for managing data, data flows and data quality techniques.



Yours is not the only organisation that is dealing with data quality.

Prioritisation

To make the data of highest value to the greatest number of users in the shortest possible time, it may be necessary to prioritise the capture and/or validation of the data (see also comments under *Completeness*, below). In order to do this, it may be necessary to:

- focus on the most critical data first,
- concentrate on discrete units (taxonomic, geographic, etc.),
- prioritise on type specimens and important vouchers
- ignore data that are not used or for which data quality cannot be guaranteed (i.e. records with poor geo-referencing information – but bear in mind the importance of some poorly georeferenced historic data),
- consider data that are of broadest value, are of greatest benefit to the majority of users and are of value to the most diverse of uses,

- work on those areas whereby lots of data can be cleaned at lowest cost (e.g. through use of batch processing).



*Not all data are created equal, so focus on the most important, and if data cleaning is required, make sure it **never** has to be repeated.*

Completeness

Organisations should strive for completeness of data (or of discrete units of the data through prioritisation – e.g. for a taxonomic category, a region, etc.) so that all eligible records are used in compiling the data. It is better to complete the data for a discrete unit and make that available, than have lots of incomplete data available as analyses carried out on incomplete data will not be comprehensive. It is also important to have a missing data policy that defines missing data thresholds and corresponding responses, along with a policy of documenting the completeness of the data (see under *Documentation*, below).

Currency and Timeliness

There are three key factors related to the timeliness or currency of data:

- Over what period were the data collected?
- When were the data last updated to reflect changes in the real world?
- How long are the data likely to remain current?

Data currency is an issue often raised by users. Many data custodians tend to use currency to refer to the period when the data were originally collected or surveyed. Because of the delay between collection and publication (which with biological data can be an exceedingly long time), the published information is a representation of “what was” and not of “what is”. Most users of biodiversity data are aware of this and that forms one of the values of these types of data, and is what makes them quite different from most other data types.

In data quality management terms, currency is more often used in the context of a “use-by” period for the data (sometimes also called timeliness), and could be related to when the data were last checked and/or updated. This may be especially relevant with respect to the names attached to the data. When were these last updated, and do they accord with the latest taxonomy? Where modern taxonomic rules of nomenclature are followed, if a species is split into a number of smaller taxa, one of those smaller taxa retains the name of the broad concept. It can be important to a user to know whether the name used refers to the broad or the narrow concept. Currency may be used as an equivalent to a “use-by” date similar to that used for food products, beyond which the custodian doesn’t guarantee the nomenclatural information attached to the record.

It may also be the case that for many datasets timeliness and currency may not be relevant or possible to include or maintain. This may apply to large museum or herbarium collections, for example. On the other hand, it may be important for observation or survey data where vouchers may not exist, or where updates have not been made to the data following recent taxonomic revisions. It is also an important issue for secondary collections, including collections that have been combined by an external agency from a number of contributing agencies. An example may be where a number of developing country institutions make their data available to a hosting institution for provision to the GBIF portal and which is not presented live from the database.

Update frequency

The frequency of update of the data within a dataset is related to currency and timeliness and needs to be formalised and documented. This includes the addition of new data as well as the frequency of release of corrected data. Both of these have an affect on the quality of the data, and are thus important for users. A user does not want to go to the expense of downloading or obtaining a dataset if it is just about to be updated and improved.

Consistency

Redman (1996) recognized two aspects of consistency: *Semantic consistency* - where the view of the data should be clear, unambiguous and consistent; and *structural consistency*, in which entity types and attributes should have the same basic structure and format. A simple example of semantic consistency is where the data are always in the same fields, and thus are easy to find – for example there are separate fields for infraspecific rank and infraspecies name so that it is always clear that the infraspecies name field includes just a name or epithet (see Table 1) and is not mixed so that sometimes it includes just a name, and at other places include a prefix of “var.” or “subsp.” followed by the name, etc. (see Table 2)

Genus	Species	Infraspecies
Eucalyptus	globulus	subsp. bicostata
Eucalyptus	globulus	bicostata

Table 1. Showing semantic inconsistency in the *Infraspecies* field.

Genus	Species	Infrasp_rank	Infraspecies
Eucalyptus	globulus	subsp.	bicostata
Eucalyptus	globulus		bicostata

Table 2. Showing semantic consistency in the *Infraspecies* field by addition of a second (“*Infrasp_rank*”) field.

Good design of a relational database would not permit many of these issues to occur, however, many existing databases used by collections’ institutions are not so well designed.

Structural consistency occurs where there is consistency within a field, for example the “*Infrasp_rank*” field (Table 2) would always have subspecies recorded the same way – not sometimes as “subsp.”, others as “ssp.”, “subspecies”, “subspec.”, “sspecies”, etc. This can be avoided through good design of the database with well structured attributes.

Consistency in both methods and documentation is important as it allows users to know what tests have been carried out and how, where to find the information, and how to interpret important pieces of information. Consistency, however, needs to be balanced against flexibility (Redman 2001).

Flexibility

Data custodians need to retain flexibility in their data quality control methods, as although much biological data are similar in nature, different approaches to data quality may be appropriate with data from different regions (for example, what associated datasets are available to check the data against), different taxonomic groups (aquatic versus terrestrial organisms, etc.), or different methods of data capture (observational or survey records versus vouchered museum collections, etc.).

Taxonomic opinions are in reality hypotheses, and differing (valid) taxonomic opinions (hypotheses) can lead to the same organism being classified differently by different taxonomists and thus having one or more alternative names – each of which may be equality valid (Pullan *et al.* 2000, Knapp *et al.* 2004). An example is where two taxonomists disagree

as to the placement of taxa within different genera – for example, some taxonomists place certain species in the genus *Eucalyptus*, whereas others believe they belong in the genus *Corymbia*. In practice, and especially in zoology, the view of the most recent reviser is accepted unless there is good reason to reject that view.

Flexibility allows the capacity for a view to change in order to accommodate new or different demands. Recent work by the Taxonomic Databases Working Group (TDWG)² and others has focused on database structures that allow for presentation of these alternate concepts (Berendsohn 1997) and, although on the surface flexibility of this nature may appear to reduce the quality, in reality it allows users greater flexibility in determining fitness for use and in those cases may thus be increasing the perceived quality.

Transparency

Transparency is important because it improves confidence in the assessment by those using the data. Transparency means making sure that errors are not hidden, but are identified and reported, that validation and quality control procedures are documented and made available, and that feedback mechanisms are open and encouraged.

An example where transparency is important is in the documentation of collection methodologies (especially important with observational and survey data). Again, this assists the user in being able to determine if the data are suitable for their particular use.

Performance measures and targets


Performance measures are a valuable addition to quality control procedures, and ensure that individual data users can be confident in the level of accuracy or quality in the data.

Performance measures may include statistical checks on the data (for example, 95% of all records are within 1,000 meters of their reported position), on the level of quality control (for example – 65% of all records have been checked by a qualified taxonomist within the previous 5 years; 90% have been checked by a qualified taxonomist within the previous 10 years), completeness (all 10-minute grid squares have been sampled), etc., etc.

Performance measures help quantify the data quality. Advantages are that:

- the organisation can assure itself that certain data are of (documented) high quality,
- they assist in overall data management and in reducing redundancy, and
- they help coordinate the various aspects of the data quality chain so that they can be organised to be carried out by different operators.

Data Cleaning



Before measuring data quality levels, first consider how users of the results might use them and then structure the results so that they can be used most effectively.

The principles of data cleaning will be covered in the associated document *Principles and Methods of Data Cleaning*. Suffice to say that a general framework for data cleaning as modified from Maletic and Marcus (2000) is:

- Define and determine error types
- Search and identify error instances
- Correct the errors

² <http://www.tdwg.org/>

- Document error instances and error types
- Modify data entry procedures to reduce incidence of similar errors in future.



Don't be seduced by the apparent simplicity of data cleaning tools. They are valuable and help in the short-term but, over the longer-term, there is no substitute for error prevention.

Outliers

The detection of outliers (geographic, statistical and environmental) can provide one of the most useful tests for finding possible errors in spatial data. It is important, however, that validation tests do not uncritically delete data because they are found to be statistical outliers. Environmental data are notorious for records that appear to be outliers statistically but which are perfectly good records. This may be due to historical evolutionary patterns, changing climate regimes, a remnant following human activities, etc. The uncritical exclusion of outliers can remove valuable records from the data set and skew future analyses.

Users, on the other hand, may decide to delete outliers from their analysis if they are unsure of their validity as good records. The identification of outliers thus not only assists data custodians to identify possible errors, but can aid users in determining whether individual data records are fit for use in their analysis or not.



Outlier detection can be a valuable validation method, but not all outliers are errors.

Setting targets for improvement

The setting of simple, easy to quantify targets can lead to a rapid improvement in data quality. A target such as to cut the percentage of new poorly-geocoded records in half every six months for two years can lead to total cut in the error rate of 94% (Redman 2001). Such targets should focus on:

- clear and aggressive time frames,
- rates of improvement rather than actual quality values,
- clear definitions (such as for 'poorly geocoded'),
- targets that are simple and achievable.

Longer term targets may also be introduced along the lines of reducing the (non value-added) time required for data cleaning by half every year by improving data entry and validation techniques.



Performance targets are a good way for an organisation to maintain a consistent level of quality checking and validation – for example 95% of all records are documented and validated within 6 months of receipt.

Auditability

It is important for custodians to know what data have been checked and when. This helps redundancy and stops data records falling through the cracks and being missed. The best way of doing this is to maintain a documented audit trail of validation.

Edit controls

Edit controls involve business rules that determine the permitted values for a particular field. For example, the value in the month field must be between 1 and 12, the value for day must be between 1 and 31 with the maximum value also dependent upon the month etc. Univariate rules apply to a single field (e.g. the month example, above), bivariate rules apply to two fields (e.g. the combination of day and month).

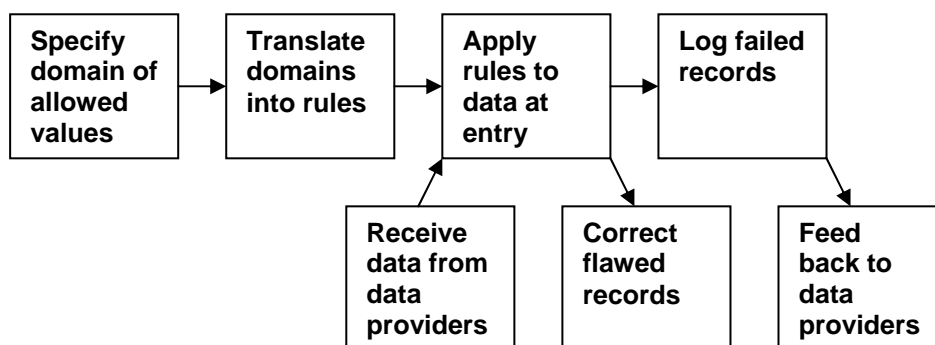


Fig. 5. Use of edit controls (modified from Redman 2001).

A second example is with coordinate data. Simple range tests will test (if the data are in latitudes and longitudes) that the latitude is between 0 and 90 degrees, minutes and seconds are between 0 and 60, etc. Once one moves to UTM data, however, it gets more complicated. Quite often a database that includes data from a small region that falls into one UTM Zone will not include the Zone within the database. This may appear to be quite acceptable as long as the data are never combined with data from other regions. But once an attempt to combine the data is made, the data becomes quite unusable. Thus editing controls need to ensure that the appropriate Zone is always included.

Minimise duplication and reworking of data

Experience in the business world has shown that the use of information management chain (see figure 3) can reduce duplication and re-working of data and lead to a reduction of error rates by up to 50% and reduce costs resulting from the use of poor data by up to two thirds (Redman 2001). This is largely due to efficiency gains through assigning clear responsibilities for data management and quality control, minimising bottlenecks and queue times, minimising duplication through different staff re-doing quality control checks, and improving the identification of better and improved methods of working.

Maintenance of original (or verbatim) data

It is important that the original data as recorded by the collector, or even inserted later by curators, etc., not be lost in the editing and data cleaning process. Changes to the database made during the data cleaning process should be added as additional information, with the original information also maintained. Once information has been deleted, it is difficult or even impossible to recover. This can be particularly important with collector and location information. What appears to a later curator as an error may not be an actual error. Changes from one location name to another (e.g. from Czechoslovakia to the Czech Republic, for example), changes not just the name, but also the circumscription. It may be important later, to know what was originally written and not just have the “corrected” version. See also comments under Archiving.

Categorisation can lead to loss of data and quality

Categorisation of data can often lead to a loss of data and thus to a reduction in overall data quality. An example may be with the collection of data with detailed locality information (and possibly even geo-referencing) but then storing the data on a grid cell basis. It is nearly always better to store the data at their finest resolution, and then categorize them on output if that is required for a particular use. If a user needs to produce a presence/absence map on a 10 X 10 minute grid, then that is easy to do from data stored as points, but if the data is stored in the database in grid cells, it is impossible to do anything with the data on a finer scale. It also makes it extremely difficult (and maybe even impossible) to combine data that may have been categorized using a different grid scale or origin. The same is the case with descriptive data – if the data is categorized into states that may be needed for a key (e.g. > 6m = tree; < 6m, = shrub), and new data is obtained from another source that used 4m instead of 6m for their tree definition, then what do you do with those between 4 and 6 meters. It is far better to store the data in exact meters, and worry about whether it is a tree or shrub later.

One case where this often occurs is with the storage of geocode accuracy. I have always recommended storing geocode accuracy in meters, but a lot of databases store this information in categories (<10m, 10-100m, 100-1000m, 1000-10,000m). If you have a record that you have been able to determine is accurate to 2km, then you have immediately lost information by having to place it into the 10km accuracy category.

Documentation

Good documentation is a key principle of data management. Without good documentation, the user cannot determine the fitness of the data for the use they have in mind and hence cannot determine the quality of the data for the purpose. A more detailed discussion on documentation is given under *Documentation*, below.

Feedback

It is essential that data custodians encourage feedback from users of their data, and take the feedback that they receive seriously. As mentioned under *User responsibility*, above, the user often has a far better chance of picking up certain error types through combining data from a range of sources, than does each individual data custodian working in isolation.

The development of good feedback mechanisms is not always an easy task. A feedback button can be placed on the query interface page, or an attachment sent to users at the time of downloading data setting out methods for feeding back data errors and comments to the custodians. Some of these are expanded upon in the associated paper on *Principles and Methods of Data Cleaning*.



Effective feedback channels with users and suppliers is an easy and productive mechanism of improving data quality.

Education and training

Education and training at all levels of the information chain can lead to vastly improved data quality (Huang *et al.* 1999). This starts with the training and education of collectors in the use of good collection procedures and implementation of the needs of the data users, through training of data input operators and technical staff responsible for the day to day management of the databases, through to education of final users as to the nature of the data, its limitations and potential uses. The education and training aspects of data quality are largely dependent on good documentation.

An example of the integration of data quality checks, education and training can be seen in the MaPSTeDI geo-referencing project (University of Colorado 2003). The process involves checking a certain number of each geocode operator's records. With a new operator, the first 200 records are checked for accuracy by a supervisor. Not only does this maintain the quality of the data, it allows the operator to learn and improve from making mistakes. Depending on the operator, an additional 100 records may be checked and as the operator becomes more experienced, checking is reduced to a random selection of 10 % of records and eventually to around 5%. If a high percentage of errors are still being discovered, then additional records are checked.

Well-designed procedures such as these can assist in educating the new user. Conversely if there are no procedures, there is little way of ensuring consistency between operators and between tasks.

Accountability

The assigning of accountability for overall data quality can assist organisations to achieve a consistent level of quality control, provide a point of reference for feedback on errors, and provide a point of contact for documentation and queries.



Poor training lies at the root of many data quality problems.

Taxonomic and Nomenclatural Data

Poor taxonomic data can “contaminate” related areas of studies (Dalcin 2004).

Taxonomy is the theory and practice of classifying organisms (Mayr and Ashlock 1991). Most of the species data we are considering here include a taxonomic (or nomenclatural) portion (i.e. the name of the organism and its classification) - termed the “Classification data domain” by Dalcin (2004). The quality of this part of the data and how the quality may be determined differs considerably from the spatial part of the data, as it is usually more abstract and more difficult to quantify.

The taxonomic data consist of (not all are always present):

- Name (scientific, common, hierarchy, rank)
- Nomenclatural status (synonym, accepted, typification)
- Reference (author, place and date of publication)
- Determination (by whom and when the record was identified)
- Quality fields (accuracy of determination, qualifiers)

One of the major sources of errors in taxonomic names is that of misspellings. Detecting spelling errors in taxonomic database can be a straightforward task when it involves scientific names that represent taxonomic hierarchies such as Family and Genus names (Dalcin 2004). In these cases standard authority files are generally available for most taxonomic groups. Increasingly, also, comprehensive lists of species names are becoming available through such projects as Species 2000 (<http://www.species2000.org>) and the ECat work program of GBIF (<http://www.gbif.org/prog/ecat>). The use of species names or epithets alone without their associated genus as an authority file is seldom satisfactory as many specific epithets may have minor variations in the name from one genus to another. One method for spelling error checking is to detect and isolate errors in scientific names, using similarity algorithms in order to identify a pair of scientific names which have a high degree of similarity but are not exactly the same (Dalcin 2004, CRIA 2005).

By far the most satisfactory method of reducing the likelihood of spelling errors in scientific names is to build authority files into the database entry process using pick lists of genus and species names, family names, etc. In an ideal situation where authority files are available, the use of these techniques should reduce the incidence of this type of error to practically zero. Unfortunately, there are large areas of the world, and a number of major taxonomic groups for which such lists are still unavailable.

Where authority files are imported from an external source such as the Catalogue of Life or ECat, then the Source-Id should be recorded in the database so that changes that are made between editions of the authority source can be easily incorporated into the database, and the database updated. Hopefully, before long this may become easier through the use of Globally Unique Identifiers (GUIDs)³.

The taxonomic quality of the data relies heavily on the available taxonomic expertise. The Taxonomic Impediment (Environment Australia 1998) and the worldwide decline in adequately trained research taxonomists will lead to a decrease in the long-term quality of production taxonomy and in the resultant quality of primary species data (Stribling *et al.* 2003). The Global Taxonomic Initiative (GTI) (CBD 2004) is attempting to remove or ameliorate the so-called “taxonomic impediment”, but the problem is likely to continue to be an issue well into the future. The quality may also decay with time, especially in cases where

³ <http://www.webopedia.com/TERM/G/GUID.html>

vouchered specimens are not available or maintained (for example with most observational data and a majority of survey data) or in those areas where relevant taxonomic expertise is not available.

The capacity of an institution to produce high quality taxonomic products (including documented primary species data) is influenced by (after Stribling *et al.* 2003):

- the level of training and experience of staff,
- the level of access to technical literature, reference and voucher collections and taxonomic specialists,
- the possession of appropriate laboratory equipment and facilities, and
- access to the internet and the resources available there.

Recording of accuracy of identification etc.

Traditionally, museums and herbaria have had a determinavit system in operation whereby experts working in taxonomic groups from time to time examine the specimens and determine their circumscription or identification. This is often carried out as part of a revisionary study, or by an expert who happens to be visiting an institution and checks the collections while there. This is a proven method, but one that is time-consuming, and largely haphazard. There is unlikely to be anyway around this, however, as automated computer identification is unlikely to be an option in the near or even long-term.

One option may be the incorporation of a field in databases that provides an indication of the certainty of the identification. The date of determination is usually incorporated in most collection databases. Such an option would be composed of a code field, and may be along the lines of (Chapman 2004):

- identified by World expert in the taxa with high certainty
- identified by World expert in the taxa with reasonable certainty
- identified by World expert in the taxa with some doubts
- identified by regional expert in the taxa with high certainty
- identified by regional expert in the taxa with reasonable certainty
- identified by regional expert in the taxa with some doubts
- identified by non-expert in the taxa with high certainty
- identified by non-expert in the taxa with reasonable certainty
- identified by non-expert in the taxa with some doubt
- identified by the collector with high certainty
- identified by the collector with reasonable certainty
- identified by the collector with some doubt.

How one might rank these would be open to some discussion, and likewise whether these were the best categories or not. I understand that there are some institutions that do have a field of this nature, but at this stage, I have not been able to find an example. The HISPID Standard Version 4 (Conn 2000) does include a simplified version – the Verification Level Flag with five codes, viz:

0	The name of the record has not been checked by any authority
1	The name of the record determined by comparison with other named plants

2	The name of the record determined by a taxonomist or by other competent persons using herbarium and/or library and/or documented living material
3	The name of the plant determined by taxonomist engaged in systematic revision of the group
4	The record is part of type gathering or propagated from type material by asexual methods

Table 3. *Verification Level Flag in HISPID (Conn 2000).*

Many institutions already have a form of certainty recording with the use of terms such as: “aff.”, “cf.”, “*s. lat.*”, “*s. str.*”, “?”. Although some of these (aff., cf.) have strict definitions, their use by individuals can vary considerably. The use of *sensu stricto* and *sensu lato* imply variations in the taxonomic concept.

In addition, where names are derived from other than taxonomic expertise, one could list the source of the names used (after Wiley 1981):

- descriptions of new taxa
- taxonomic revisions
- classifications
- taxonomic keys
- faunistic or floristic studies
- atlases
- catalogues
- checklists
- handbooks
- taxonomic scholarship/rules of nomenclature
- phylogenetic analysis

Uncertainty can usually be reduced, and quality improved through comparison of two or more publications or specialists. Differences between identifications between taxonomists, however, may not necessarily imply that one of the identifications is an error, but may show a difference in taxonomic opinion (i.e. differing hypotheses) as to the placement of the taxon.

Precision of identification

According to Stribling *et al.* (2003), identification precision (which they wrongly termed taxonomic precision) may be evaluated by comparing the results of a randomly selected sample that is processed by two taxonomists or specialists. An assessment may also be made by comparing the names given to duplicate specimens held (and identified) by different institutions. These are fairly abstract notions, and I am not sure of the value in recording this type of information.

A second part to identification precision, however, is the level to which a specimen is identified. An identification to species, or subspecies, is a more precise identification than one to just family or genus. In documenting a dataset, it may be of value to users to know that 50% of identifications are to genus only – a case with many faunal groups.

Bias

Bias is systematic error that arises from a uniform shift in values (Chrisman 1991). It often arises from a consistently applied methodology that leads to error that is systematic in nature. Bias in taxonomic nomenclature can arise where the identification is precise, but not accurate. Such bias might arise from the misinterpretation of a dichotomous key or

morphological structure, the use of an invalid nomenclature or outdated publication (Stribling *et al.* 2003) or the use of an inappropriate publication (e.g. a flora of another area to that being studied and that may not have all the relevant taxa from the area being studied).

Consistency

Inconsistency can occur within the classification domain in databases if two or more names are considered as “accepted” and to represent the same taxon (eg. *Eucalyptus eremaea* and *Corymbia eremaea*). This may relate to differing opinions as to the taxonomy, or errors due to alternate spellings (for example, *Tabernaemontana hystrix*, *Tabernaemontana histryx* and *Tabernaemontana histrix* – CRIA 2005).

Completeness

Motro and Rakov (1998 from Dalcin 2004) referred to completeness as “*whether all the data are available*” and divided data completeness into the completeness of files (no records are missing), and the completeness of records (all fields are known for each record).

Completeness in taxonomic terms (i.e. with a names or taxon database) refers to the coverage of names. Does the database include names at all levels in the hierarchy (e.g. down to subspecies or only species)? What portion of the animal or plant kingdom does the database cover? Does the database include synonyms? All of these are important in assisting the user to determine the fitness of the data for his/her particular use. Dalcin (2004), for example, divides completeness into *nomenclatural completeness* representing inclusion of all possible names, given a context, (e.g. in a taxonomic context - a list of all names for a specific taxonomic group; or in a spatial context - a list of all names for a specific region) and *classification completeness* representing all possible names related to an “accepted” name for a given taxon (i.e., a complete synonymy).

With a specimen or observational database, completeness may be along the lines “are all Darwin Core fields included” and “do all Darwin Core fields include data”. In a character database, “are characters for all necessary life-stages present” (e.g. fruits of plants, instars of insects).

Voucher collections

The importance of voucher collections cannot be over stressed, however it is not always possible for databases to include vouchers. Many observational databases are made without at the same time making voucher collections. It is also not possible for political, legal, conservation or other purposes to take a sample for vouchering in all cases or areas.

Where vouchering is possible it is often a valuable exercise at the initial stages of species-based programs to develop cooperative agreements between data collectors and institutions such as museums or herbaria to support the deposition of reference and voucher collections (Brigham 1998). Such agreements should also cover appropriate archiving and disposal strategies, including minimum time periods before disposal or archiving.

Spatial Data

Spatial data has often led the field in the development of standards for data documentation (for example with the development of the Spatial Data Transfer Standards (USGS 2004), the ISPIRE (Information for Spatial Information in Europe) program⁴ and many more) and has since been at the forefront of the development of data quality standards (e.g. ISO 19115 for Geographic Information – Metadata⁵). The numerical nature of much of the spatial data means that they are more open to the use of statistical procedures than the taxonomic data, and have thus allowed the development of a number of data quality checking methods (see accompanying paper on *Principles and Methods of Data Cleaning*).

This does not mean that all spatial parts of the data (the “Field data domain” of Dalcin 2004) are easy to digitise or are accurate. Many historical collections in museums and herbaria have only very basic textual descriptions of the localities of collections, and it is a major effort to convert these to numerical geocodes or coordinates. This can be exacerbated by the nature of many of these collections, for example, collected at a time when detailed maps were not available to collectors, and where many of the locality names used no longer occur in published gazetteers or maps. To add geo-referencing information to historical records, especially where good historical gazetteers don’t exist, can be quite time-consuming and result in quite low levels of accuracy.

A number of tools have been developed to assist users to georeference their data, including on-line tools and guidelines. These will be expanded on in the associated paper on *Principles and Methods of Data Cleaning*. In addition, most collectors are now using GPS (Global Positioning Systems) to record geocodes at the time of collection. For a discussion on the accuracies associated with the use of GPS see the chapter on “*Capturing Data*”

The testing of errors in already assigned georeferences can involve:

- checking against other information internal to the record itself or between records within the database - for example, State, named district, etc.;
- checking against an external reference using a database – is the record consistent with the collecting localities of the collector?
- checking against an external reference using a GIS – does the record fall on land rather than at sea?
- checking for outliers in geographic space; or
- checking outliers in environmental space.

All of these methods will be expanded upon in the accompanying paper on *Principles and Methods of Data Cleaning*.

Spatial Accuracy

How is the positional accuracy of spatial data measured?

For most GIS layers (topographic maps, etc.) the source of ‘truth’ is relatively easy to determine as there are usually external sources of higher accuracy of a few features in the database – survey trig points, street and road intersections, etc. (Chrisman 1991). Many of the tests, though, are not simple and documentation – such as the US National Map Accuracy Standard – complicated. Traditionally, spatial accuracy is determined by comparison to a number of “well-defined” points along with specified acceptable levels of error, measured as

⁴ <http://www.ec-gis.org/inspire/>

⁵ <http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35>

root-mean-square deviation (RMSE) from zero to determine accuracy (Chrisman 1991). RMSE is not easy to apply to individual points, however, and is more applicable to whole datasets or digital maps. With individual points the distance from the true location using a simple point-radius method (Wieczorek *et al.* 2004) or similar methods are simple and easy to use. There are two factors involved – how accurately the well-defined point can be determined will determine the accuracy of the point being tested, and the accuracy and precision of measurement of the tested point will add to the error. For example, if the road intersection can only be accurately determined to within 100 meters, the centroid of the collection point is then a 100-meter circle before the precision of that point is added (see comments in Wieczorek 2001).

The US Federal Geographic Data Committee (FGDC) released the Geospatial Positioning Accuracy Standards (GPAS) in 1998. These standards include separate sections for Geodetic Networks and for Spatial Data Accuracy (FGDC 1998).

- *‘The NSSDA uses root-mean-square error (RMSE) to estimate positional accuracy. RMSE is the square root of the average of the set of squared differences between dataset coordinate values and coordinate values from an independent source of higher accuracy for identical points.’*
- *‘Accuracy is reported in ground distances at the 95% confidence level. Accuracy reported at the 95% confidence level means that 95% of the positions in the dataset will have an error with respect to true ground position that is equal to or smaller than the reported accuracy value. The reported accuracy value reflects all uncertainties, including those introduced by geodetic control coordinates, compilation, and final computation of ground coordinate values in the product.’*

Examples of map accuracy statements made in Australia using such methods with regards to their products are:

- *‘The average accuracy of this map ± 100 meters in the horizontal position of well defined detail and ± 20 meters in elevation.’* (Division of National Mapping, Sheet SD52-14, Edition 1, 1:250,000).

These accuracies need to be added to any determination of the geo-referencing of a collection based on a paper or digital map. As there is always uncertainty in spatial data accuracy, no absolute statement of accuracy can be applied and it is important that known accuracy be documented. Errors are propagated throughout the information chain and contribute to uncertainties in the final results, whether it is a map product from a GIS or a species model using distributional modelling software (Heuvelink 1998).

BioGeomancer project

A project⁶ has recently been funded by the Gordon and Betty Moore Foundation to assist in improving the geo-referencing of primary species records and in assessing, improving and documenting accuracy. This project should report and make available the developed tools sometime during 2006.

False Precision and Accuracy

An additional factor to be aware of is that of False Precision and Accuracy. Many GIS users are unaware of the issues involved in spatial data accuracy, error and uncertainty, and often assume that their data are absolute. They often report levels of accuracy that are unattainable

⁶ <http://www.biogeomancer.org/>

with their source data. Many institutions are now using GIS to help with their geo-referencing and by zooming in to levels not supported by the data (and using decimal degrees) can end up with a precision that is unrealistic. Also, with the use of a GPS to record the location of a collection event, location is often reported to 1 or 2 meters when in reality many hand-held GPS units being used are probably only accurate to around 10 meters or less. This is particularly relevant with using a GPS to determine altitude (see comments under *Capturing Data* below).

Collector and Collection Data

Information on the collector and the collection (the Collection data domain of Dalcin 2004) includes information about the collection itself – the collector, date of collection and additional information such as habitat, soils, weather conditions, observers experience, etc. They may be categorised as (modified from Conn 1996, 2000)

- Collection author(s) and collector's number(s)
- Observers experience, etc.
- Collection date/period(s)
- Collection method (particularly for observation/survey data)
- Associated data

Many of these issues will vary considerably with the type of data being collected – be it for a museum collection, an observation or results of a detailed survey. With a static collection such as that for a museum, the collector's name and number, and date are key attributes, along with associated data such as habit, habitat, etc., and maybe capture method (for animals). For observational data, such things as length of observation, area covered by observation, time of day (start and end times in addition to date), and associated data such as weather conditions, sex of observed animal, activity, etc. With survey data, information on the survey method, size (grid and total area), effort, weather conditions, frequency, whether vouchers were collected and their numbers, etc. along with many of those mentioned for observations.

Attribute Accuracy

Issues that may impinge on data quality with respect to the collection information, include the way collectors' names, numbers, initials etc. are recorded (Koch 2003), the accuracy of date and time recording, the consistency in recording of associated data at time of collection, such as habit, habitat, soils, vegetation type, flower colour, sex, associated species.

An example of a problem that regularly arises with collection data is “collector's number” – where some collectors don't use unique numbers to identify their collections. This can cause a loss of quality as these tags are sometimes used to help identify locations of collections, identifications, duplicate collections in different institutions, etc.

Consistency

Consistency in use of terminology with respect to the collection domain is often quite erratic, and it is rare that associated data fields, in particular, are consistent within a dataset, let alone across different datasets.

Completeness

Completeness of collection information is also usually quite variable. More often than not, habitat, collector's number, flowering etc. will not be completed for many records. This makes a study of habitat, for example, difficult from just collections alone.

Descriptive Data

Descriptive databases are increasing being used to both store data and as a method of publication, often in place of traditional publications. Morphological, physiological and phenological data elements are examples of data in this domain. Descriptive data are often used to generate information for use in cladistic analysis and automatically generated descriptions and identification tools.

The Taxonomic Databases Working Group (TDWG) has had a long history in the development and promotion of standards in the area of descriptive databases – firstly with its support of the DELTA standard (Dallwitz and Paine 1986) and more recently with the development of the “Structure of Descriptive Data” working group (<http://160.45.63.11/Projects/TDWG-SDD/>).

Quality of descriptive data can be variable, and although the data elements are often measured, in reality the accuracy may be determined by cases where the data are unobservable (e.g. with historical data), impractical to observe (e.g. too costly) and/or perceived rather than real (e.g. subjective evaluation such as colour, abundance, etc.).

In most cases, descriptive data are stored at the species level rather than at the specimen level and is thus usually averaged or ranged. As pointed out by Morse (1974 as reported by Dalcin 2004), taxonomic information is inherently of a lower degree of reliability than specimen observation data. Irrespective of this, there is a greater tendency in recent times to store, at least some of these data, at the specimen level with a resultant increase in quality.

Completeness

At the specimen level, completeness of recordings of descriptive data may depend on the quality of the specimen, time of year etc. For example, it may not be possible to record fruit or flower characteristics from the same specimen. For this reason, many fields will of necessity be left blank. In other cases, the attribute may not be relevant to the character and thus not all attributes will be scored.

Consistency

Inconsistency issues can arise between two related data items. For example, two species descriptor characteristics may be scored as (Dalcin 2004):

- “HABIT=HERBACEUS” and
- “USES=WOOD”

Inconsistent representation of the same attribute may also affect quality, especially where poor attribute definitions are used or consistent standards are not rigidly adhered to. For example (Dalcin 2004):

- “FLOWER COLOUR= CARMINE”, and
- “FLOWER COLOUR=CRIMSON”.

The use of standard terminologies can help reduce the degree of error and mis-interpretation considerably. Standard terminologies are being developed in a range of areas and disciplines,

and the recent move to the development of federated descriptive databases has increased the consistency with which terminologies are used. The development of the TDWG Standard for the Structure of Descriptive Data (SDD) (TDWG 2005) can only assist this process.

Capturing Data

There are a variety of ways to capture primary species data and species-occurrence data, each having its own levels of precision and accuracy, as well as their own sources of error and uncertainty. Each of these have differing impacts on the final “fitness for use”, or quality, of the data. Several of the more common methods used with species data are briefly discussed.

Opportunistic

A majority of species-occurrence data have been collected opportunistically. Many of these records are now stored as specimens in museums and herbaria. Most of the historic data included only a textual location reference such as 5 km NW of a town, etc. and were seldom given a georeference at the time of collection. The addition of a georeference has usually been carried out at a later date, and usually by someone other than the collector (Chapman and Busby 1994). Many observational records (bird atlas data, etc.) have also been collected opportunistically.

These data are usually captured digitally often in batch format, and the geo-referencing generally done by reference to physical maps. They usually include both significantly low precision and accuracy. The majority of these data cannot be regarded as being of greater accuracy than about 2-10 km.

Field Survey

Field survey data have generally included a spatial reference, often in the form of latitude and longitude or a UTM reference. The spatial reference can usually be regarded as having an accuracy of about 100 –250 meter accuracy. Care must be taken, however, as to what the spatial reference is referring to – it may not be the location of the actual observation, but may refer, for example, to the mid point of a transect, or the corner (or centre) of a grid square, and this is not always clear. In addition, as records are seldom vouchered (i.e. a physical collection made and stored for later reference), the taxonomic accuracy cannot always be relied upon. This is particularly so the further one moves away from the time of the survey, and as taxonomic concepts alter.

Broad-scale Observations

Some biological surveys may only record data within a particular boundary or grid cell. For example, a survey of the species within a National Park, or bird observations made within 10-minute grid squares (e.g. Birds Australia 2001, 2003). The accuracy of records such as these may only be in the order of 1-10 km or greater.

Global Positioning Systems (GPS)

Global Positioning Systems, or GPSs have increasingly come into play with the collection of species data. These include not only survey data, but also opportunistic and observational collections.

GPS technology uses triangulation to determine the location of a position on the earth’s surface. The distance measured is the range between the GPS receiver and the GPS Satellites

(Van Sickle 1996). As the GPS satellites are at known locations in space, the position on earth can be calculated. A minimum of four GPS satellites is required to determine the location of a position on the earth's surface (McElroy *et al.* 1998, Van Sickle 1996). This is not generally a limitation today, as one can often receive 7 or more satellites in most locations on earth, however historically, the number of satellites receivable was not always sufficient. Prior to May 2000, most GPS units use by civilians involved "Selective availability". Its removal has greatly improved the accuracy that can generally be expected (NOAA 2002).

Before removal of Selective availability, the accuracy of *Hand-held GPS* receivers as used by most biologists and observers in the field was in the order of about 100 meters or worse (McElroy *et al.* 1998, Van Sickle, 1996, Leick 1995). Since then, however, the accuracy of GPS receivers has improved and today, most manufacturers of hand-held GPS units promise errors of less than 10 meters in open areas when using 4 or more satellites. The accuracy can be improved by averaging the results of multiple observations at a single location (McElroy *et al.* 1998), and some modern GPS receivers that include averaging algorithms can bring the accuracy down to around 5 meters or maybe even better.

The use of *Differential GPS* (DGPS) can improve the accuracy considerably. DGPS uses referencing to a GPS Base Station (usually a survey control point) at a known location to calibrate the receiving GPS. This works through the Base Station and hand-held GPS referencing the satellites' positions at the same time and thus reduces error due to atmospheric conditions. In this way the hand-held GPS applies the appropriate corrections to the determined position. Depending on the quality of the receivers being used, one can expect an accuracy of between 1 and 5 meters. This accuracy decreases as the distance of the receiver from the Base Station increases. Again averaging can further improve on these figures (McElroy *et al.* 1998).

The Wide Area Augmentation System (WAAS) is a GPS-based navigation and landing system developed for precision guidance of aircraft (Federal Aviation Administration 2004). WAAS involves ground-based antennae whose precisely known locations can provide greater positional accuracy with the use of GPSs. Similar technologies such as Local Area Augmentation System (LAAS) are also being developed to provide even finer precision.

Even greater accuracies can be received using either *Real-time Differential GPS* (McElroy *et al.* 1998) or *Static GPS* (McElroy *et al.* 1998, Van Sickle 1996). *Static GPS* uses high precision instruments and specialist techniques and is generally only used by surveyors. Surveys conducted in Australia using these techniques reported accuracies in the centimetre range. These techniques are unlikely to be extensively used with biological record collection due to the cost and general lack of requirement for such precision.

To obtain accuracies such as those reported above, the GPS Receiver must be located in an area that is free from overhead obstructions and reflective surfaces and have a good field of view to the horizon (for example, they do not work very well under a heavy forest canopy). The GPS receiver must be able to record signals from at least four GPS satellites in a suitable geometric arrangement. The best arrangement is to have "*one satellite directly overhead and the other three equally spaced around the horizon*" (McElroy *et al.* 1998). The GPS Receiver must also be set to an appropriate datum for the area, and the datum used recorded.

GPS Height. Most biologists know little about the height determined using a GPS. It is important to note that the height displayed by a GPS receiver is actually the height in relation to the Earth Centric Datum (and is thus related to the surface of the earth's ellipsoid) and not a height that relates to Mean Sea Level or to a standard height datum such as the Australian

Height Datum. In Australia, for example, the difference between the height reported from a GPS receiver and Mean Sea Level can vary from –35 to +80 meters and tends to vary in an unpredictable manner (McElroy *et al.* 1998, Van Sickle 1996).

Data Entry and Acquisition (Capturing data electronically)

Data entry and acquisition is inherently prone to errors both simple and complex. (Maletic and Marcus 2000)

Basic data capture

The first step in data capture is usually the capture of information from a specimen label, journal, field notebook, accession book or card catalogue. This may be done through use of skilled or un-skilled data entry operators or through electronic scanning of information. The level of error due to data entry can often be reduced through double-keying, using learning and training software associated with scanning, and through using experts and supervisors to carry out testing of entry on a sample-basis (see the MaPSTeDI Guidelines mentioned below).

User-interfaces

The development of a specific data-entry User Interface can also be a way of decreasing data-entry errors. Many institutions use unskilled staff or volunteers as data-entry operators and the development of a simple (non-technical) user interface that data entry operators feel comfortable with can increase the accuracy of entry. Such an interface can help data input by being able to quickly search authority fields, existing entries in the database, other related databases, and even use search engines such as Google that can help an operator decide on the correct spelling or terminology where they may have difficulty reading a label, or determining what should and shouldn't go into particular fields. In some cases this can be applied through database design that incorporates Authorities tables and drop-down menus (pick lists) that precludes unskilled data-input personnel having to make decisions about names, localities, or habitats.

Geo-referencing

Maps are one of the most effective ways of communicating information and this alone justifies the recent increase in databasing and georeferencing of specimen data from museums and herbaria, along with the increase in capture of georeferenced observational information. The enhanced data handling ability of maps allows us to better study, identify, visualize, document and correct errors and uncertainties (Spear *et al.* 1996). It also provides a powerful method for visualizing and communicating the uncertainties inherent in the data, and thus be able to present users with a way of determining the quality, or fitness of use of the data.

Capturing data electronically and attaching geocodes (i.e. geo-referencing the data) can be a difficult and time-consuming task. Results from the MaPSTeDI project (University of Colorado 2003) suggest that a competent operator can georeference one record every 5 minutes. Other studies (Armstrong 1992, Wieczorek 2002) have shown that geo-referencing may take significantly longer – for example, the MANIS database suggests a rate of about 9 per hour for US, 6 per hour for non-US North American, and 3 per hour for non-North American localities (Wieczorek 2002).

MaNIS/HerpNet/ORNIS

Georeferencing Guidelines

<http://manisnet.org/manis/GeorefGuide.html>

MaPSTeDI

Georeferencing in MaPSTeDI

<http://mapstedi.colorado.edu/geo-referencing.html>

A
nu

number of excellent methods and guidelines have been developed to assist data managers with geo-referencing. The Georeferencing Guidelines developed by John Wieczorek at the Museum of Vertebrate Zoology in Berkeley (Wieczorek 2001) and the MaPSTeDI (Mountains and Plains Spatio-Temporal Database Informatics Initiative) guidelines (University of Colorado 2003) are two of the most comprehensive studies of the topic conducted to date and I refer the reader to those Guidelines. The guidelines cover the determination of the accuracy and precision of a point derived from a textual locality, uncertainties arising from the use of different datums, effects of using different map scales, etc. They are comprehensive coverages of the topic and I would hope that readers of this document might regard them as integral adjuncts to this document.

There are also a number of on-line tools that can assist with the determination of geocodes – for example for places at a given distance and direction from a known locality. These will be covered in more detail in the associated document on *Principles and Methods of Data Cleaning*.



(Peabody Museum of Natural History)

<http://www.biogeomancer.org/>



(Reference Centre for Environmental Information)

<http://splink.cria.org.br/tools/>

Error

Tools such as those mentioned earlier are powerful tools for reducing error and increasing quality. But no geocoding method can totally eliminate error. As stated in the MaPSTeDI Guidelines:

“While geocoding is not an exact science and no collection can be geocoded 100% correctly, quality checking can drastically improve the percentage of the collection that is correctly geocoded. Every project should take it into account when planning their geocoding operation” (University of Colorado 2003).

One common source of georeferencing error is through the uncritical use of electronic gazetteers. In some cases these gazetteers have been developed from projects for publishing hardcopy maps, and the location of the gazetted point given in the gazetteer is the bottom left hand corner of where the name was to be written on the map, and not

the location of the point to which it referred (e.g. The Australian Gazetteer prior to 1998 developed by the Australian Land Information Group). Hopefully, most gazetteers have been corrected, but there may already be georeferences added to museum and herbarium data based on these values. The accuracy of such records should be checked by means of random spot checks of localities against gazetteers or accurate large scale maps.



It is often quicker and more efficient to carry out geo-referencing as a separate activity following the digitisation of the label information. This allows the database to be used to sort collections by locality, collector, date, etc. and allows for more efficient use of maps for obtaining geocode information. It also saves duplication of geocoding of multiple records from the same locality, etc.

Documenting Data

“Metadata is data about data. It is a description of the characteristics of data that has been collected for a specific purpose.” (ANZLIC 1996a).

Good documentation occurs at both the dataset level and at the data record level.

Metadata provides information about datasets such as content, extent, accessibility, currency, completeness, fitness-for-purpose and suitability-for-use. When metadata are provided, a user can gain an understanding of the quality of a dataset and ascertain the suitability of that dataset prior to using it. Good metadata allows for improved data exchange, searching and retrieval. Metadata usually refers to the whole dataset, however some see documentation of data at the record level (such as the recording of accuracy) as record-level metadata.

Irrespective of what it may be called, good documentation at both the dataset level and at the record level is important.

All data include error – there is no escaping it! It is knowing what the error is that is important, and knowing if the error is within acceptable limits for the purpose to which the data are to be put. This is where metadata come to the fore for datasets as a whole, and indeed it is in the area of metadata development that the term “fitness for use” has come to prominence. The concept of fitness for use did not become fully recognised as an important one with spatial information until the early nineties, and it wasn’t until the mid 90s that it started to appear in the literature in this context (Agumya and Hunter 1996).

Recording information only at the dataset level, however, will not always supply the information that the user requires. Recording error at the record level, especially with species data, can be extremely important for determining the fitness of that record for use. When this information is available, a user can request, for example, only those data that are better than a certain metric value – e.g. better than 5,000 meters. It is also important that automated geo-referencing tools include calculated accuracy as a field in the output.

It is also important that users of the data understand the concept of fitness for use. All too often species-occurrence data are extracted from a database in a “record no., x, y” format regardless of any accuracy information that may be present. The coordinate itself is always represented as a point, but it seldom, if ever, refers to a true point. Some records may have been entered into a database with an arbitrary point (for example a collection that just has “South America” on the label), and given an accuracy of 5 000 000 meters in the accuracy field. There are some databases that do this! To extract the record and use its arbitrary point



The data must be documented with sufficient detailed metadata to enable its use by third parties without reference to the originator of the data.

will be extremely misleading. Users need to be made aware that there is an accuracy field if it is present, and be advised on how to use it. In cases where data providers develop standard data reports, they should make it mandatory that the accuracy field be included when data are supplied.

Fig. 6. Example of search of data using the MaPSTeDI search tool <http://www.geomuse.org/mapstedi/client/textSearch.html>. The example shows the ability to search for data of a particular accuracy using documentation at the record level.

The screenshot shows the MaPSTeDI search interface with the following elements:

- Title:** Search MaPSTeDI Collections
- Search Fields:** Taxon, Common Name, Locality, State (dropdown), County (dropdown), Accuracy (dropdown), Institution (dropdown), Date from, Date to.
- Accuracy Dropdown Menu:** Exact Coordinates, Amended Coordinates, Legal Description, Within 1 KM, Within 5 KM, Within 10 KM, To the county, To the state, To the project region.
- Record Type Selection:** Radio buttons for Abridged Record and Complete Record.
- Navigation:** Links for Advanced Search and Basic Map.
- Buttons:** Search button, and a dropdown for 'on: 50'.

Documenting accuracy, precision and error in spatial data is essential if users are to be able to determine the quality of those data for their purposes. Such documentation should include (as a minimum):

- dataset title
- source of data
- data lineage (actions performed on data since their collection or derivation)
- accuracy (positional, temporal and attribute)
- logical consistency
- date and life expectancy of the data (data currency and status, frequency of update)
- data field definitions
- collection methodology
- completeness
- conditions of use and use constraints (e.g. copyright, license restrictions etc).
- custodianship and contact information

It is worth defining some of these terms as not all data custodians are familiar with them. Many of these terms refer to a collection of data in a database rather than to the individual collection records themselves.

Positional accuracy

Positional accuracy refers to how closely the coordinate descriptions of features compare to their actual location (Minnesota Planning 1999). Where possible and known, the Geodetic Datum used to determine the coordinate position should be noted.

It is also recommended, that databases include a field to record the positional accuracy of each individual record. There are a number of ways of doing this. Some databases use a code, however, it is preferred that a simple metric value be used to represent the estimated accuracy of the record (Chapman and Busby 1994, Conn 1996, 2000, Wieczorek *et al.* 2004). This can be important for users extracting data for a particular purpose – for example, they may only want data that are accurate to better than 2000 meters. Sometimes, it may also be of value to include a field at the record level on how the georeference information was determined. For example;

- use of differential GPS
- handheld GPS corrupted by Selective Availability (e.g. prior to 2002)
- A map reference at 1:100 000 and obtained by triangulation using readily identifiable features
- A map reference using dead reckoning
- A map reference obtained remotely (eg. in a helicopter)
- Obtained automatically using geo-referencing software using point-radius method.
- Use of gazetteer including name, date and version of the gazetteer.

Attribute accuracy

Attribute accuracy refers to an assessment of how correctly and reliably the features in the data are described in relation to their real world values. Ideally it should include a list of attributes and information on the accuracy of each. For example,

Records are provided by experienced observers. Additional accuracy is obtained by testing the correctness of attributes against vouchered specimens lodged at the museum or herbarium for expert verification. Approximately 40% of plant records are verified with voucher specimens, amphibians 51%, mammals 12 %, reptiles, 18% and birds 1%. (SA Dept. Env. & Planning 2002).

Lineage

Lineage refers to the sources of the data, along with the steps taken to process the dataset to bring it to its present state. It may include the collection method (i.e. “data collected in a 10 X 10 meter grid”) and information on validation tests that have been carried out on the data. The history of the processing steps may include:

- the data capture method(s)
- any intermediate processing steps and methods
- the methods used to generate the final product
- any validation steps carried out on the data.

For example;

The data were collected using 20 meter x 20 meter fixed quadrats. Total species counts, structure and other habitat data were also collected. The data were classified using Twinspan into groups comprising similar groups of species.

Logical consistency

Logical consistency provides a brief assessment of the logical relationships between items in the data. Although for most data collected here (museum and herbarium data) some of these items may not be relevant, however they may be for some observational data (check-lists of species in a National Park or bioregion, etc.) and some survey data. For spatial data where the data are stored digitally, logical consistency tests can be carried out automatically. Such things as

- Are all points, lines and polygons labelled and do any have duplicate labels?
- Do lines intersect at nodes or cross unintentionally?
- Are all polygon boundaries closed?
- Are all points, lines and polygons topologically related?

Logical consistency can also apply in the case of datasets where there are other logical relationships between items or objects in the dataset. In such cases a description of any tests carried out on the relationships should be included. Examples may be dates that occur in different fields - if the date given in one field says the project was carried out between years 'a' and 'b' but the date of recording of an attribute in another field is outside that range, then this is logically inconsistent; or records are out of geographic range - if one field records the fact that data were collected in Brazil, yet another field includes latitude and longitudes for records from Paraguay, then this is a logical inconsistency between the two fields.

Documentation of checks carried out is an important part of metadata. Checks may include tests such as "point-in-polygon" checks and is used for such purposes in the GIS world. See an expansion on methods in the associated paper on *Principles and Methods of Data Cleaning*.

Completeness

Completeness refers to both the temporal and spatial coverage of the data or dataset as a portion of the total possible extent of the data. Documentation of completeness is an essential component for determining quality. Examples may include:

Complete for areas north of 30° S, scattered records only between 30° and 40° S.

Dataset covers only records prior to 1995 collected largely opportunistically, mainly from New South Wales, but includes some records from other States.

From a user perspective, completeness relates to "all the data they need" (English 1999). That is the user needs to know if the database includes all the fields that they need for their analysis and needs to know the "completeness" of some of those fields. For example the user may want to carry out a study comparing attributes over time, but if the database only includes data up to a certain year, it may not be useable for the analysis (see second example, above).

Accessibility

For the data to be of value to a user it needs to be accessible. Not all data are available on-line and to access some data the user may need to contact the custodian and seek permission to access it, or to obtain a copy of what they need on CD. Documentation of access (and use) conditions is important for users to be able to access the data and is therefore an aspect of data quality. Documentation of accessibility may include:

- Contact addresses for the data
- Access conditions
- Access method (if data available electronically)
- Data format

- Caveats
- Copyright information,
- Costs if applicable
- Restrictions on use

Temporal accuracy

Temporal accuracy refers to the accuracy of the information in time. For example: “*data only accurate to month*”. This can be important in databases where the “day” field may not allow for a null value and in cases where the information is not available, automatically puts a “1” in the field. This can lead to a false impression of the precision. This is even more important where a record is known only to the year and the database automatically records it as the 1st of January. If a user is studying the flowering period of plants or migratory patterns of birds, for example, then they need to know this information so that they can exclude these records as (for their purpose) the data quality is low, and is not “fit for use”.

Documenting validation procedures

One of the keys to knowing what error exists in data is documentation. It is of very little use to anyone if checks of data quality are carried out, and corrections made, if it is not fully documented. This is especially important where these checks are being carried out by other than the originator of the data. There is always the possibility that perceived errors are not errors at all, and that changes that are made, add new error. It is also important that checking not be done over and over again. We cannot afford to waste resources in this way. For example, data quality checks carried out on data by a user may identify a number of suspect records. These records may then be checked and found to be perfectly good records and genuine outliers. If this information is not documented in the record, further down the line, someone else may come along and carry out more data quality checks that again identify the same records as suspect. This person may then exclude the records from their analysis, or spend more valuable time rechecking the information. This is basic risk management, and should be carried out routinely by all data custodians and users. The value and need for good documentation cannot be stressed too heavily. It assists users in knowing what the data are, what the quality is, and what purposes the data are likely to be fit for. It also aids curators and data custodians to keep track of the data and their quality and to not waste resources rechecking supposed errors.

Documentation and database design

One of the ways of making sure that error is fully documented is to include it in the early planning stages of database design and construction. Additional data quality/accuracy fields can then be incorporated. Fields such as positional or geocode accuracy, source of information for the georeference information and elevation, fields for who added the information – was the coordinate data added by the collector using a GPS, or a data entry operator at a later date using a map at a particular scale, was the elevation automatically generated from a DEM, if so, what was the source of the DEM, its date and scale, etc. All this information will be valuable in later determining whether the information is of value for a particular use or not, and the user of the data can then decide.

“data users need to exercise caution when basing biological assessments on taxonomic data sets that do not specifically present documentation of at least some performance characteristics”. (Stribling *et al.* 2003).

Storage of data

The storage of data can have an effect on data quality in a number of ways. Many of these are not obvious, but need to be considered both in the design of the storage vessel (database) and as a unit in the data quality chain.

The topic of selection or development of a database is too large a topic to be covered here, and should be the subject of a separate study. A study commissioned by GBIF examined Collection Management Software (Berendsohn *et al.* 2003) and I refer readers to that document.

This section examines some of the main principles of data storage as they relate to data quality.

Backup of data

The regular backup of data helps ensure consistent quality levels. It is essential that organisations maintain current disaster recovery and back-up procedures. Whenever data are lost or corrupted, there is a concomitant loss in quality.

Archiving

Archiving (including obsolescence and disposal) of data is an area of data and risk management that needs more attention. Data archiving, in particular by universities, NGOs and private individuals should be a priority data management issue. Universities have high turnovers of staff and often the research data are stored in a distributed manner – usually in the researchers own PC or filing cabinet. If not fully documented, such data can very quickly lose their usability and accessibility. More often than not it is discarded sometime after the researcher has left the organisation, as no one knows what it is or cares to put the effort in to maintaining it. It is for this reason that Universities in particular need sound documenting and archiving strategies.

Individual researchers working outside of a major institution need to ensure that their data are maintained and/or archived after their death, or after they cease to have an interest in it. Similarly NGO organisations that may not have long-term funding for the storage of data, need to enter into arrangements with appropriate organisations that do have a long-term data management strategy (including for archiving) and who may have an interest in the data.

Data archiving has become much easier in recent years with the development of the DiGIR/Darwin Core and BioCASE/ABCD⁷ protocols. These provide an easy way for an institution, University department or individual to export their database in one of these formats and to store them in XML format, either on their own site, or forwarded to a host institution. This is an easy way to store data in perpetuity and/or to make them available through distributed search procedures such as GBIF's Data Portal.

The cleanup and disposal and archiving of data are also issues with data on the World Wide Web. Web sites that are abandoned by their creators, or that contain old and obsolete data leave cyberspace littered with digital debris (various references). Organisations need a data archiving strategy built into their information management chain. The physical archiving of data is too large a topic to enter into here, however, a recent document on archiving data through the use of CDs and DVDs has been published by the Council on Information and

⁷ <http://www.tdwg.org>;
<http://www.gbif.org/links/standards>

Library Resources and the United States National Institute of Standards and Technology (Byers 2003). It is a valuable summary of this technology and readers may like to refer to it.



Data which are no longer required (for legal or other reasons) should not be destroyed, or put at risk without exploiting all other possibilities – including archiving (NLWRA 2003).

Data Integrity

Data integrity refers to the condition in which data have not been altered or destroyed in an unauthorised manner, and has not been accidentally or maliciously modified, altered, or destroyed (such as by a virus or voltage spike).

Data often change—for example, when the taxonomic information in a record is updated following a redetermination—but users expect that the computer system will maintain the integrity of the data and that the computer system itself will not inadvertently or incorrectly alter a value. *Data corruption* is when data integrity fails and an inadvertent or incorrect change occurs.



Data integrity is preserved through good data management, storage, backup and archiving.

Patterns of error

Taxonomic and species-occurrence databases, like all databases, are vulnerable to content error patterns. English (1999) recognised the following error patterns which he called data defects. Dalcin (2004) adopted these for use with taxonomic databases. The values here are from English (1999) with examples cited from Chapman (1991) and from the databases of the Australian Virtual Herbarium⁸ and Brazil's speciesLink⁹:

- **Domain value redundancy** –Non-standardised data values, or synonym values exist and in which two or more values or codes have the same meaning. Redundancy is very typical with descriptive data if standardised terminologies are not followed, or where compilation of data from different sources is badly controlled.
- **Missing Data Values** – A data field that should contain a value, doesn't. This includes both required fields and fields not required to be entered at data capture, but are needed in downstream processing. Examples include geo-referencing or coordinate values (latitude and longitude).
- **Incorrect Data Values** – These may be caused by transposition of key-strokes, entering data in the wrong place, misunderstanding of the meaning of the data captured, not being able to read the writing on the label, or where mandatory fields require a value but the data entry operator does not know a value for entry. Incorrect data values are the most obvious and common errors and can affect every data value in every field. Spelling errors in scientific names is a common pattern associated

⁸ <http://www.cpbr.gov.au/avh/>

⁹ <http://specieslink.cria.org.br/>

with incorrect data values in taxonomic and nomenclatural databases (see discussion elsewhere), and the placement of a zero in geo-referencing fields, etc.

- **Nonatomic Data Values** – Occurs when more than one fact is entered into the same field (e.g. genus, species and author in the same field, or rank and infra-specific name). This type of error is usually a result of poorly thought out database design. This type of error pattern can cause real problems for data integration.

Genus	Species	Infraspecies
Eucalyptus	globulus	subsp. bicostata
Family	Species	
Myrtaceae	Eucalyptus globulus Labill.	

Table 4. *Examples of Nonatomic data values.*

- **Domain schizophrenia** – Fields used for purposes for which they weren't designed and which end up including data of more than one nature.

Family	Genus	Species
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

Table 5. *Examples of Domain schizophrenia*

- **Duplicate Occurrences** – Multiple records that represent a single entity. The most typical cases occur with names where alternative spellings or valid nomenclatural alternatives may occur. These can lead to difficulties for users when searching for a name, or when attempting to combine data from different databases. Examples:
 - Phaius tancarvilleae
 - Phaius tankervilleae
 - Phaius tankarvilleae
 - Phaius tankervilleae
 - Phaius tankervilleae
 - Phaius tankervilleae
 - Brassicaceae/Cruciferae (exact equivalents; both of which are allowed by the International Botanical Code).
- **Inconsistent Data Values** – Occurs when data in related databases may be updated inconsistently or at different times in the two databases. For example, between the living collection and herbarium databases, or the museum collections database and the related images database.
- **Information Quality Contamination** – Results from combining accurate data with inaccurate data. For example combining data with information at the subspecies level into a database that only includes data down to the level of species.

Spatial Data

The storage of spatial data covers the location information (textual locality information) as well as coordinate information (georeferencing data) usually given as a coordinate pair (an easting and a northing). Many databases are now beginning to include parsed or atomized

location data such as nearest named place, distance and direction in addition to the free text locality description. Several projects are now underway to improve on the parsing of free text location data to create these atomized fields and to aid in the georeferencing process. The BioGeomancer project¹⁰ recently funded by the Gordon and Betty Moore Foundation is one such project.

Geo-referencing (or coordinate) information is generally entered into databases as either latitude or longitude (spherical coordinate system) or in UTM (or related) coordinates (planimetric coordinate system). A spherical coordinate system such as latitude and longitude encircle the globe and to be represented on a paper map have to be stretched in unusual ways known as projections. Spherical coordinate systems are not equal area and the distance between one degree of latitude and the next, for example can vary considerably depending on whether one is near the equator or near a pole. Planimetric coordinate systems are closer to equal area projections and can be used for measuring or making area calculations.

Many institutions are now beginning to enter data in degrees, minutes and seconds or degrees and decimal minutes (as reported by many GPS units), and having the database convert them into decimal degrees for storage. For transfer and use in a GIS it is generally best to store the data in decimal degrees as it provides for easy data transfer and provides the highest possible accuracy.

Storage of data in UTM coordinates often occurs in institutions where the data is restricted to just the one UTM Zone. It has the advantage of being area based as discussed above so each grid is a square (or rectangle) and allows for easy representation on a flat map, or for calculating distance and area. It is important, however, when storing data in UTM (or related) coordinate systems that the Zone also be stored, otherwise difficulties arise in combining data from other areas or institutions.

Decimal Degrees

The storage of decimal degrees in many databases can lead to *False Precision* as mentioned above. The precision at which data are stored (and made available) should be a consideration. The database should not allow reporting at a precision higher than the highest precision data in the database. With most biological data, this will be to about 4 decimal places (ca. 10 meters).

Datums

There are many possible geodetic datums. The Earth is not a true sphere, but an ellipsoid, and difficulties arise when trying to fit a coordinate system to the surface of that ellipsoid (Chapman *et al.* 2005). To solve this, the concept of a 'datum' was created. A datum is a set of points used to reference a position in the sphere to the ellipsoid of revolution. Historically, different reference systems were generated for different parts of the earth, and it was only with the advent of satellites that a truly global reference system or datum could be generated, as satellites were used to fix the center of the earth. The difference in a latitude and longitude position on earth using different geodetic datums can be as much as 400 meters or more (Wieczorek 2001).

Because of the difference, it is important that databases record the datum used, otherwise when data is combined the resultant error between two recordings of the same location could be quite significant.

¹⁰ <http://www.biogeomancer.org/>

Manipulation of spatial data

There are many ways that spatial data can be manipulated. Many have no effect on the accuracy of spatial data, some do. Examples of some of the methods that do affect the positional accuracy of spatial data are

Conversion of data from one format to another

Perhaps the most common data conversions carried out by those involved with the collection, storage and use of species and species-occurrence data are the conversion of geocodes from degrees/minutes/seconds to decimal degrees (DMS to DD), or from UTM coordinates to decimal degrees (UTM to DD). Others include the conversion of miles to kilometres in textual locality descriptions, the conversion of feet to meters in altitude and depth recordings, etc.

All of these are fairly simple conversions, but can lead to a false impression of the accuracy through miss-use of precision. For example a collection that gives the altitude as 250 feet (which the collector may have meant was somewhere between 200 and 300 feet) when converted to metric would be 76.2 meters (to 1 decimal place) or perhaps 76 meters if rounded. It would better to record the converted value as 80 meters and even better include an accuracy field to add perhaps 20 meters (\pm). The false use of precision can lead to what appears to be increased accuracy, but in reality is a drop in quality.

Datums and Projections

The conversion of data from one geodetic datum to another can lead to quite significant error as the conversions are not uniform (see Wieczorek 2001 for a discussion of datums and their effect on data quality). Many countries or regions are now converting most of their data to one standard for their region - either the World Geodetic Datum (WGS84), or datums that approximate this quite closely (the Australian Geographic Datum (AGD84), in Australia which varies from WGS84 by around 10cm; and the EUREF89 in Europe which varies from WGS84 by about 20cm are two examples). The conversion from one datum position to another, for example is probably not necessary if the data are only accurate to around 5 or 10 km. If you are dealing with data of about 10-100 m accuracy, however, datum shifts can be quite significant and important (in some areas up to 400m or more – Wieczorek 2001).

Similarly, where mapped data are in polygons (e.g. collections from a national park), one needs to be aware of the errors that can arise in converting from one projection to another (e.g. Albers to Geographic). Standard formulae are available to calculate the error that arises in doing such conversions, and the metadata accompanying the data should reflect that information.

Grids

Whenever data are converted from vector format into raster or grid format, accuracy and precision is lost. This is due to the size of the grid cells in the raster file that are used to approximate the vector data (Burrough and McDonnell 1998). The precision and accuracy cannot be regained by converting the data back into vector format. For a fuller discussion on the problems encountered in using and converting raster data, and of the problems of scale see Chapman *et al.* (2004).

Data Integration

Geographical datasets are difficult to integrate when there are inconsistencies between them. These inconsistencies may involve both the spatial and attribute characteristics of the data, and may necessitate the use of various, often time-consuming, corrective measures (Shepherd 1991). Inconsistencies may result from:

- Differences in recording or measurement techniques (e.g. area size and time periods in observation data), survey methods (e.g. grid size, width of transect) or data categories (e.g. different definitions of categories with categorical data).
- Errors in measurements or survey methods (e.g. errors in transcription, data recording, identifications)
- Differences in resolution (spatial, temporal or attribute)
- Vague and imprecise definitions
- Fuzziness of objects (e.g. soil or vegetation boundaries, identifications where some are to species, others to subspecies, others only to genus)
- Differences in use or interpretation of terminology and nomenclature (e.g. different taxonomies used).
- Differences in GPS settings (datum, coordinate system, etc.)

Such integration problems are greater where the data are:

- Of different types (e.g. specimen data from a museum mixed with survey and observational data)
- From different jurisdictions (e.g. where survey methodologies may be different)
- Obtained from multiple sources
- Of multiple scales
- Consists of different data types (maps, specimen, image, etc.)
- From different time periods
- Stored in different database types, media etc. (e.g. some database software do not allow for “null” values)
- Variously parsed (e.g. where one dataset includes the whole scientific name in one field, and others have it split into separate fields for genus, species)



Data integration produces higher quality results when contributing data custodians have followed and used consistent data storage standards.

Representation and Presentation

Methods should always be developed to make the most efficient use of existing data, whatever their quality. However, in order for the data to be reliable, they must also be validated or accompanied by information that indicates the level of reliability.
(Olivieri *et al.* 1995)

In their role to understand, explain, quantify and evaluate biodiversity, scientists and scientific institutions are increasingly recognised as information providers. This recognition is based on the ability to provide reliable and useable information to decision-makers, managers, the general public, and others. Ambiguous, confused, incomplete, contradictory and erroneous information, as a result of poorly managed databases, can affect their reputation as information providers and scientific authorities (Dalcin 2004).

A key purpose of digital data handling in the biological sciences is to provide users of information with a cost-effective method of querying and analysing that information. In that sense, its success is determined by the extent to which it can provide the user with an accurate view of the biological world. But the biological world is infinitely complex and must be generalised, approximated and abstracted to be represented and understood (Goodchild *et al.* 1991). Ways of doing this are through the use of geographic information systems, environmental modelling tools and decision support systems. In using these tools, however, it is essential that variation be sampled and measured, and error and uncertainty be described and visualised. It is in this area that we still have a long way to go to reach what could be regarded as best practice.

Biology was one of the first disciplines to develop techniques for error reporting with the use of error bars and various statistical measures and estimates. The reporting of error was not seen as a weakness because error estimates provide crucial information for correct interpretation of the data (Chrisman 1991). In the delivery of species data, similar error



Effective data quality programs help prevent embarrassment to the organisation and individuals – both internally and publicly.

reporting techniques

need to be developed and used, so that users of these data have similar abilities to correctly interpret and use the data.

Determining Users' Needs

Determining users' needs is not a simple process, and it is difficult to develop detailed requirements and then structure the data to meet those requirements. But it is important to identify key users and to work with them to develop their needs and requirements. Good data-user requirements can lead to better and more efficient data collection, data management and overall data quality.

Relevancy

Relevancy is closely related to "quality" and refers to the relevancy of data for the use required of it. It may relate to something as simple as trying to use a Flora for an area for which it wasn't intended, but for which nothing else exists, or to data that may be in a different projection than that required and which may require considerable work to make it useful and "relevant".

Believability

Believability is the extent to which data are regarded by the user as being credible (Dalcin 2004). It is often subject to the user's perception or assessment of the data's suitability for their purpose and may be based on previous experience or a comparison to commonly accepted standards (Pipino *et al.* 2002). The reputation of a dataset can sometimes depend upon the perceived believability (and thus useability) of users, but it is something that can often be improved upon by good documentation.

Wang et al. (1995) include a diagram that relates many of these topics into a hierarchical representation and shows the relationship between entities such as believability and reputation, etc.

Living with uncertainty in spatial data

Uncertainty, especially with spatial data, is a fact of life, but often uncertainty in the data has not been well documented, and is not always obvious to users. The proliferation of easy to use desktop mapping systems has allowed non GIS-professionals to easily visualize and analyse spatial relationships in their data, but this is often done using inappropriate scales (Chapman *et al.* 2005), and without regard to the spatial error and uncertainty inherent in the data (Chapman 1999). In some instances this can lead to a dangerous misuse of the data, and occasionally to tragic consequences (Redman 2001). Recently there has been an increase in simple online map services that allow users to view and analyse spatial data as in a traditional desktop GIS but allows the publisher of the service to control the data layers and the scale of the data sets that appear. In the near future this will expand even further with the development of functional Web Mapping Services (WMS). The control of data layers and scale by the publishers of the map (e.g. allowing different layers to be turned on or off automatically as the user zooms in) reduces some of the simple mistakes that otherwise could be made.

It is essential that uncertainty in data be documented, firstly through the use of good metadata, and secondly through visualisation and presentation. One area of research that needs pursuing with respect to species and species-occurrence data is the development of techniques to visualize uncertainty – for example to show footprints of accuracy. Instead of a collection record being represented as a point of latitude and longitude there is a need to include the accuracy associated with the record and thus present the location as a footprint – a circle, an ellipse, etc., and maybe even include levels of probability (Chapman 2002).

It is important that those that know the data and their limitations with regard to positional and/or attribute accuracy assist users by documenting and making available that information in order to guide users in determining the fitness of the data for their use.

Visualisation of error and uncertainty

There is still a long way to go to develop good error visualisation methods for species data, although a number of new and exciting methods are being developed (e.g. Zhang and Goodchild 2002). Perhaps the easiest methods are through the use of an error layer as an extra overlay in a GIS. Such techniques have been used in the cartographic world where a layer may provide shading of different intensities to show the reliability of different parts of the map. Other techniques could involve the use of different symbols (a dotted line as opposed to a solid line, dots of different size or intensity, etc. to indicate data of lower quality or accuracy). The use of such overlays often may also provide clues as to the origin of the errors and these can be a valuable tool in the validation and checking of data.

The use of a misclassification matrix whereby rows provide expected results, and columns observed results, is useful where such statistical calculations are possible. In these cases errors along rows are errors of omissions and errors along columns errors of commission (Chrisman 1991). Such methods do not generally lend themselves to use with species-occurrence data, but may be of value, for example, with survey data where presence/absence records are observed over a period of time.

Risk Assessment

Decision makers would prefer a climate of certainty; however natural systems are inherently variable and seldom conform to this desire. Risk assessment techniques are increasingly providing decision makers and environmental managers with estimates of certainty and risk, so that environmental decisions can be made with greater certainty. In the case of species, where knowledge of their exact occurrence is often scant, areas of 'likely occurrence' may be used as a surrogate. Within broad areas of 'likely occurrence', however, there may be areas that are more 'likely' than others (Chapman 2002).

The concept of risk can generally be seen as having two elements – the likelihood and magnitude of something happening and the consequences if and when an event does happen (Beer and Ziolkowski 1995). In a species data context, risk assessment may extend from the risk of an on-site fire destroying data if off-site backup procedures are not implemented through to the risk of an environmental decision being in error due to use of poor quality data. An example of this may be the cost involved in prohibiting a development because of information that a threatened species occurs in the area. In some environmental situations, governments are increasingly looking at applying the *precautionary principle* in making important environmental decisions.

Legal and moral responsibilities

There are a number of areas where legal and moral responsibilities may arise with respect to the quality and presentation of species data. These include

- Copyright and Intellectual Property Rights;
- Privacy;
- Truth in Labelling;
- Restricted presentation of quality for sensitive taxa;
- Indigenous Rights;
- Liability;
- Caveats and disclaimers

In most cases the *Copyright and Intellectual Property Rights* in the data can be covered by documentation accompanying the data. Where these may vary from record to record, then it should be recorded at the record level, otherwise it can be covered in the metadata.

A number of countries have recently introduced *privacy* legislation, and data custodians should be aware of the provisions of such legislation. This can be particularly relevant where data are being transferred across political boundaries or made available via the Internet. In some countries, information about individuals cannot be stored in a database or made available without their express permission. How this may affect information attached to species-occurrence data is not clear, however, custodians should be aware of the issue and make provision for it where necessary.

Good quality control measures along with good metadata will usually lead to compliance with "*truth in labelling*" concepts. So far, in legislation at least, "truth in labelling" has been largely restricted to food products. It is however mentioned in papers dealing with the development of a Global Spatial Data Infrastructure (Nebert and Lance 2001, Lance 2001), National Spatial Data Infrastructure for the USA (Nebert 1999) and an Australian and New Zealand Spatial Data Infrastructure (ANZLIC 1996b). In the Global SDI paper (Lance 2001), it is recommended that a Spatial Data Clearinghouse should include "*a free advertising*

method to provide world access to holdings under the principle of ‘truth-in-labeling’”, and to quote from the Australian and New Zealand document:

“Land and geographic data quality standards may be descriptive, prescriptive, or both. A descriptive standard is based on the concept of ‘truth in labelling’, requiring data producers to report what is known about the quality of the data. This enables data users to make an informed judgement about the ‘fitness for purpose’ of the data.”

Restricted presentation of quality with sensitive species may occur where locality information is “fuzzed” - for example to restrict knowledge of the exact location of threatened species, trade sensitive species, etc. This is a reduction in the published quality of the data, and where this does occur it should be clearly documented so that users know what they are getting, and can decide if the data are then of value for their use or not.

Indigenous rights may also affect the data quality, as there may be cases where some information has to be restricted due to sensitivities of indigenous peoples. Documentation to the effect that “some data have been restricted for purposes of complying with the rights of indigenous peoples” should then be included.

In 1998, Epstein *et al.* examined the issue of legal liability in relation to the use of spatial information. Some key points that they make are:

- *There is now ‘considerable potential’ for litigation and for loss of both personal and organisational reputation and integrity arising from error in spatial information.*
- *Traditional disclaimers may not be a strong defence in the event of litigation.*
- *In order to limit liability, organisations may be required to maintain a high level of quality documentation that adequately and truthfully labels their products to the ‘best of their ability and knowledge’.*

Caveats and disclaimers are an important part of the documentation of data quality. They should be written in a way as to not only cover the custodian organisation, but to also supply the user with some idea as to the quality of the data, and what may be able to be expected from that quality.



Most agencies and groups involved in producing data will be judged on the ease at which the data and information is made available, and the quality of the information. Those that are able to publish, share, access, integrate and use information are those that will benefit most (NLWRA 2003).

Certification and Accreditation

Can and should species-occurrence data be certified? With increased data becoming available from many agencies, users want to know which institutions they can rely on, and which follow documented quality control procedures. Should they just rely on well known institutions, or are there lesser-known institutions also with reliable data? What data available from the better-known institutions are reliable and which aren't. *Reputation* alone can be the deciding factor on where a user may source their data but reputation is a subjective concept and is a fragile character on which to base actions and decisions (Dalcin 2004). Is this what we want in our discipline? Good metadata and documentation of data quality procedures can often turn a subjective factor such as reputation into something that users can base a more scientific and reasoned assessment on. Perhaps we should develop a certification and

accreditation process that informs users of organisations that conform to minimum data quality documentation standards and procedures.

The development of agreed quality certification could lead to an improvement in overall data quality and to increased certainty among users on the value of the data. This in-turn could lead to improved funding for certified organisations. Dalcin (2004) suggests that “*a quality certification of taxonomic data could involve three aspects: primary data sources (the raw material), the information chain (the process) and the database (the product).*”

Peer Review of databases

A peer review system for databases could be introduced for species databases. Such a peer review process could feed into a certification procedure as examined above, and may involve issues such as quality control procedures, documentation and metadata, update and feedback mechanisms, etc.

Conclusion

One goal of any information specialist is to avoid needless error. By directly recognizing error, it may be possible to confine it to acceptable limits. Still error cannot always be avoided cheaply or easily.

(Chrisman 1991).

The importance of data quality and error checking cannot be stressed too strongly. As stressed throughout this document, it is essential if the data are to be of real value in developing outputs that will lead to improved environmental decisions and management. Data quality is an important issue with all data, be they museum or herbarium collection data, observational records, survey data, or species check lists. There is a merging requirement by many governments around the world for data to be of high quality and be better documented. For example:

- There is a strong direction from the Australian Federal, State and Territory Governments to improve services and make more effective use of resources, including data and information resources.
- There is an increasing recognition that data collected at public expense must be properly managed in order to make it accessible to the public so as to realise its potential and justify the considerable production and maintenance costs involved.
- There is increasing pressure from customers for easier and quicker access to the right data and information and that they are provided at little or no cost.
- There is an increased focus within governments for the need to rationalise and combine data in order to improve efficiency and add value.
- There is an increasing requirement that data be relevant. This applies to new collections, new surveys, to data management and publication.

The need for quality data is not in question, but many data managers assume that the data contained and portrayed in their system is absolute and error free – or that the errors are not important. But error and uncertainty are inherent in all data, and all errors affect the final uses that the data may be put to. The processes of acquiring and managing data to improve its quality are essential parts of data management. All parts of the information quality chain need to be examined and improved by organisations responsible for species-occurrence data and their documentation is a key to users being able to know and understanding the data and to be able to determine their “fitness for use” and thus their quality.

The human factor is potentially the greatest threat to the accuracy and reliability of spatial information. It is also the one factor that can ensure both the reliability, and generate an understanding, of the weaknesses inherent in any given spatial data set (Bannerman 1999).

References

- Agumya, A. and Hunter, G.J. 1996. Assessing Fitness for Use of Spatial Information: Information Utilisation and Decision Uncertainty. *Proceedings of the GIS/LIS '96 Conference*, Denver, Colorado, pp. 359-70
- ANZLIC. 1996a. *ANZLIC Guidelines: Core Metadata Elements Version 1, Metadata for high level land and geographic data directories in Australia and New Zealand*. ANZLIC Working Group on Metadata, Australia and New Zealand Land Information Council. <http://www.anzlic.org.au/metaelem.htm>. [Accessed 14 Jul 2004]
- ANZLIC 1996b *Spatial Data Infrastructure for Australia and New Zealand. Discussion Paper*. www.anzlic.org.au/get/2374268456. [Accessed 1 Jul 2004].
- Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* 5(1): 80-88.
- Bannerman, B.S., 1999. *Positional Accuracy, Error and Uncertainty in Spatial Information*. Australia: Geoinnovations Pty Ltd. <http://www.geoinnovations.com.au/posacc/patoc.htm> [Accessed 14 Jul 2004].
- Beer, T. & Ziolkowski, F. (1995). *Environmental risk assessment: an Australian perspective*. Supervising Scientist Report 102. Canberra: Commonwealth of Australia. <http://www.deh.gov.au/ssd/publications/ssr/102.html> [Accessed 14 Jul 2004]
- Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283-309.
- Berendsohn, W., Güntsch, A. and Röpert, D. (2003). Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Members/irc/gbif/digit/library?l=/digitization_collections/contract_2003_report/ [Accessed 16 Mar. 2005].
- Birds Australia. 2001. *Atlas of Australian Birds. Search Methods*. Melbourne: Birds Australia. <http://www.birdsaustralia.com.au/atlas/search.html> [Accessed 30 Jun 2004].
- Birds Australia. 2003. *Integrating Biodiversity into Regional Planning – The Wimmera Catchment Management Authority Pilot Project*. Canberra Environment Australia. <http://www.deh.gov.au/biodiversity/publications/wimmera/methods.html>. [Accessed 30 Jun 2004].
- Brigham, A.R. 1998. Biodiversity Value of federal Collections **in** Opportunities for Federally Associated Collections. San Diego, CA, Nov 18-20, 1998.
- Burrough, P.A., McDonnell R.A. 1998. *Principals of Geographical Information Systems*: Oxford University Press.
- Byers, F.R. 2003. *Care and Handling of CDs and DVDs. A Guide for Librarians and Archivists*. Washington, DC: National Institute of Standards and Technology and Council on Library and Information Resources. <http://www.itl.nist.gov/div895/carefordisc/CDandDVDCareandHandlingGuide.pdf> [Accessed 30 Jun 2004].
- CBD. 2004. *Global Taxonomic Initiative Background*. Convention on Biological Diversity. <http://www.biodiv.org/programmes/cross-cutting/taxonomy/default.asp> [Accessed 13 Jul 2004].

- Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 **in** Lowell, K. and Jatón, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.
- Chapman, A.D. 2002. Risk assessment and uncertainty in mapped and modelled distributions of threatened species in Australia pp 31-40 **in** Hunter, G. & Lowell, K. (eds) *Accuracy 2002 – Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Melbourne: Melbourne University.
- Chapman, A.D. 2004. Environmental Data Quality – b. Data Cleaning Tools. Appendix I to *Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota. FAPESP/Biota process no. 2001/02175-5 March 2003 – March 2004*. Campinas, Brazil: CRIA 57 pp. http://splink.cria.org.br/docs/appendix_i.pdf [Accessed 14 Jul. 2004]
- Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 **in** Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.
- Chapman, A.D., Muñoz, M.E. de S. and Koch, I. 2005. Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context. *Biodiversity Informatics* **2**: 24-41.
- Chrisman, N.R. 1983. The role of quality information in the long-term functioning of a GIS. *Proceedings of AUTOCART06*, 2: 303-321. Falls Church, VA: ASPRS.
- Chrisman, N.R., 1991. The Error Component in Spatial Data. pp. 165-174 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
- Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data*. Version 3. Sydney: Royal Botanic Gardens.
- Conn, B.J. (ed.) 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data*. Version 4 – Internet only version. Sydney: Royal Botanic Gardens. <http://plantnet.rbg Syd.nsw.gov.au/Hispid4/> [Accessed 30 Jun. 2004].
- Cullen, A.C. and Frey, H.C. 1999. *Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York: Plenum Press, 335 pages.
- CRIA 2005. *speciesLink. Dados e ferramentas – Data Cleaning*. Campinas, Brazil: Centro de Referência em Informação Ambiental. <http://splink.cria.org.br/dc/> [Accessed 4 Apr. 2005].
- Dalcin, E.C. 2004. Data Quality Concepts and Techniques Applied to Taxonomic Databases. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf [Accessed 7 Jan. 2004].
- Dallwitz, M.J. and Paine, T.A. 1986. *Users guide to the DELTA system*. CSIRO Division of Entomology Report No. 13, pp. 3-6. *TDWG Standard*. <http://biodiversity.uno.edu/delta/> [Accessed 9 Jul 2004].
- Davis R.E., Foote, F.S., Anderson, J.M., Mikhail, E.M. 1981. *Surveying: Theory and Practice*, Sixth Edition: McGraw-Hill.
- DeMers M.N. 1997. *Fundamentals of Geographic Information Systems*. John Wiley and Sons Inc.

- English, L.P. 1999. *Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits*. New York: John Wiley & Sons, Inc. 518pp.
- Environment Australia. 1998. *The Darwin Declaration*. Canberra: Australian Biological Resources Study. <http://www.biodiv.org/programmes/cross-cutting/taxonomy/darwin-declaration.asp> [Accessed 14 Jul 2004].
- Epstein, E.F., Hunter, G.J. and Agumya, A.. 1998, Liability Insurance and the Use of Geographical Information: *International Journal of Geographical Information Science* 12(3): 203-214.
- Federal Aviation Administration. 2004. Wide Area Augmentation System. <http://gps.faa.gov/Programs/WAAS/waas.htm> [Accessed 15 Sep. 2004].
- FGDC. 1998. *Geospatial Positioning Accuracy Standards*. US Federal Geographic Data Committee. http://www.fgdc.gov/standards/status/sub1_3.html [Accessed 14 Jul. 2004].
- Foote, K.E. and Huebner, D.J. 1995. *The Geographer's Craft Project*, Department of Geography, University of Texas. <http://www.colorado.edu/geography/gcraft/contents.html> [Accessed 14 Jul 2004].
- Gad, S.C. and Taulbee, S.M. 1996. *Handbook of data recording, maintenance, and management for the biomedical sciences*. Boca Raton: CRC Press.
- Goodchild, M.F., Rhind, D.W. and Maguire, D.J. 1991. *Introduction* pp. 3-7 In: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
- Heuvelink, G.B.M. 1998. *Error Propagation in Environmental Modeling with GIS*: Taylor and Francis.
- Huang, K.-T., Yang, W.L. and Wang, R.Y. 1999. *Quality Information and Knowledge*. New Jersey: Prentice Hall.
- Juran, J.M. 1964. *Managerial Breakthrough*. New York: McGraw-Hill.
- Knapp, S., Lamas, G., Lughadha, E.N. and Novarino, G. 2004. Stability or stasis in the names of organisms: the evolving codes of nomenclature. *Phil. Trans: Biol. Sci.* 359(1444): 611-622.
- Koch, I. (2003). *Coletores de plantas brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 26 Jan. 2004].
- Lance, K. 2001. Discussion of Pertinent Issues. pp. 5-14 in *Proceedings USGS/EROS Data Center Kenya SCI Workshop, November 12 2001*. http://kism.icconnect.co.ke/NSDI/proceedings_kenya_NSDI.PDF [Accessed 1 Jul 2004].
- Leick, A. 1995. *GPS Satellite Surveying*: John Wiley and Sons, Inc: New York.
- Library of Congress. 2004. *Program for Cooperative Cataloging*. Washington, DC. US Library of Congress. <http://www.loc.gov/catdir/pcc/> [Accessed 26 Jun 2004].
- Lunetta, R.S. and Lyon, J.G. (eds). 2004. *Remote Sensing and GIS Accuracy*. Boca Raton, FL, USA: CRC Press.
- Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis pp. 200-209 in *Proceedings of the Conference on Information Quality (IQ2000)*. Boston: Massachusetts Institute of Technology. <http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf> [Accessed 21 November 2003].
- Mayr, E. and Ashlock, P.D. 1991. *Principles of systematic zoology*. New York: McGraw-Hill.
- McElroy, S., Robins, I., Jones, G. and Kinlyside, D. 1998. *Exploring GPS, A GPS Users Guide*: The Global Positioning System Consortium.

- Minnesota Planning. 1999. *Positional Accuracy Handbook. Using the National Standard for Spatial data Accuracy to measure and report geographic data quality*. Minnesota Planning: Land Management Information Center.
http://www.mnplan.state.mn.us/pdf/1999/lmic/nssda_o.pdf [Accessed 14 Jul. 2004]
- Morse, L.E. 1974. Computer programs for specimen identification, key construction and description printing using taxonomic data matrices. *Publs. Mich. St. Univ. Mus., biol. ser.* 5, 1–128.
- Motro, A. and Rakov, I. 1998. Estimating the Quality of Databases. *FQAS 1998*: 298-307
- Naumann, F. 2001. *From Database to Information Systems – Information Quality Makes the Difference*. IBM Almaden Research Center. 17 pp.
- Nebert, D. and Lance, K. 2001. Spatial Data Infrastructure – Concepts and Components. *Proceedings JICA Workshop on Application of Geospatial Information and GIS. 19 March 2001, Kenya*. <http://kism.iconnect.co.ke/JICAWorkshop/pdf/Ottichilo.pdf> [Accessed 1 Jul 2004].
- Nebert, D. 1999. *NSDI and Gazetteer Data*. Presented at the Digital Gazetteer Information Exchange Workshop, Oct 13-14, 1999. Transcribed and edited from audiotape.
http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/session3/nebert.htm [Accessed 1 Jul 2004].
- NLWRA. 2003. *Natural Resources Information Management Toolkit*. Canberra: National Land and Water Resources Audit. <http://www.nlwra.gov.au/toolkit/contents.html> [Accessed 7 Jul 2004].
- NOAA. 2002. Removal of GPS Selective Availability (SA).
http://www.ngs.noaa.gov/FGCS/info/sans_SA/ [Accessed 15 Sep 2004].
- Olivieri, S., Harrison, J. and Busby, J.R. 1995. Data and Information Management and Communication. pp. 607–670 in Heywood, V.H. (ed.) *Global Biodiversity Assessment*. London: Cambridge University Press. 1140pp.
- Pipino, L.L., Lee, Y.W. and Wang, R.Y. 2002. Data Quality Assessment. *Communications of ACM* 45(4): 211-218.
- Pullan, M.R., Watson, M.F., Kennedy, J.B., Raguenaud, C., Hyam, R. 2000. The Prometheus Taxonomic Model: a practical approach to representing multiple classifications. *Taxon* 49: 55-75.
- Redman, T.C. 1996. *Data Quality for the Information Age*. Artech House, Inc.
- Redman, T.C. 2001. *Data Quality: The Field Guide*. Boston, MA: Digital Press.
- SA Dept Env. & Planning. 2002. *Opportunistic Biological Records (OPPORTUNE)*. South Australian Department of Environment and Heritage.
<http://www.asdd.sa.gov.au/asdd/ANZSA1022000008.html> [Accessed 14 Jul. 2004].
- SEC 2002. *Final Data Quality Assurance Guidelines*. United States Securities and Exchange Commission. <http://www.sec.gov/about/dataqualityguide.htm> [Accessed 26 Jun 2004].
- Shepherd, I.D.H. 1991. Information Integration and GIS. pp. 337-360 in: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
- Spear, M., J.Hall and R.Wadsworth. 1996. *Communication of Uncertainty in Spatial Data to Policy Makers* in Mowrer, H.T., Czaplowski, R.L. and Hamre, R.H. (eds) *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*, May 21-23, 1996. Fort Collins, Colorado. USDA Forest Service Technical Report RM-GTR-277.

- Stribling, J.B., Moulton, S.R. II and Lester, G.T. 2003. Determining the quality of taxonomic data. *J. N. Amer. Benthol. Soc.* **22(4)**: 621-631.
- Strong, D.M., Lee, Y.W. and Wang, R.W. 1997. Data quality in context. *Communications of ACM* 40(5): 103-110.
- Taulbee, S.M. 1996. *Implementing data quality systems in biomedical records* pp. 47-75 in Gad, S.C. and Taulbee, S.M. Handbook of data recording, maintenance, and management for the biomedical sciences. Boca Raton: CRC Press.
- TDWG. 2005. TDWG Working Group: Structure of Descriptive Data (SDD). Taxonomic Databases Working Group (TDWG). <http://160.45.63.11/Projects/TDWG-SDD/> [Accessed 4 Apr. 2005].
- University of Colorado. 2003. MaPSTeDI. *Georeferencing in MaPSTeDI*. Denver, CO: University of Colorado. <http://mapstedi.colorado.edu/georeferencing.html> [Accessed 30 Jun. 2004].
- USGS. 2004. *What is SDTS?* Washington: USGS. <http://mcmcweb.er.usgs.gov/sdts/whatsdts.html> [Accessed 30 Jun. 2004].
- Van Sickel, J. 1996. *GPS for Land Surveyors*: Ann Arbor Press, Inc: New York.
- Wang, R.Y. 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM* 41(2): 58-65.
- Wang, R.Y., Storey, V.C., Firth, C.P., 1995. A frame-work for analysis of data quality research, *IEEE Transactions on Knowledge and Data Engineering* 7: 4, 623-640.
- Wieczorek, J. 2001. *MaNIS: Georeferencing Geo-referencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS <http://manisnet.org/manis/GeorefGuide.html> [Accessed 26 Jan. 2004].
- Wieczorek, J. 2002. *Summary of the MaNIS Meeting. American Society of Mammalogists, McNeese State University, Lake Charles, LA, June 16, 2002*. Berkeley: University of California, Berkeley - MaNIS. <http://manisnet.org/manis/ASM2002.html> [Accessed 30 Jun. 2004].
- Wieczorek, J., Guo, Q. and Hijmans, R.J. (2004). *The point-radius method for georeferencing locality descriptions and calculating associated uncertainty*. *International Journal for GIS* 18(8): 754-767.
- Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.
- Zhang, J. and Goodchild, M.F. 2002. *Uncertainty in Geographic Information*. London: Taylor and Francis.

Index

- accessibility, 37
- accountability, 20
- Accreditation, 48
- accuracy, 3
 - attribute, 28, 36
 - documentation of, 35
 - false, 26
 - positional, 25, 35
 - recording of
 - taxonomic data, 22
 - spatial, 25
 - temporal, 37
- archiving, 39
- attribute accuracy, 28, 36
- audit trail, 17
- bias, 23
- BioGeomancer, 26
- caveats and disclaimers, 48
- Certification, 48
- Classification data domain, 21
- collection data, 28
- Collection data domain, 28
- collector
 - responsibility of, 11
- completeness, 14, 24, 28, 29, 37
- consistency, 15, 23, 28, 29
 - semantic, 15
 - structural, 15
- copyright, 47
- data
 - archiving, 39
 - backup of, 39
 - believability, 46
 - capture, 30, 32
 - categorization of, 18
 - collection, 28
 - collector, 28
 - consistency, 28, 29
 - descriptive, 29
 - documentation of, 19
 - entry, 32
 - grid, 43
 - integration, 43
 - integrity, 40
 - nomenclatural, 21
 - observational, 30
 - opportunistic, 30
 - presentation, 45
 - relevancy, 45
 - representation, 45
 - spatial, 25, 41
 - storage, 39
 - survey, 30
 - taxonomic, 21
 - uncertainty, 46
- data cleaning, 16
- data currency, 14
- data custodian, 12
- data management, 18
- data quality
 - policy, 8
 - principles, 1
 - strategy, 9
 - vision, 8
- data user
 - definition, 7
 - responsibility of, 12
- databases
 - peer review of, 49
- decimal degrees, 42
- DELTA standard, 29
- descriptive data, 29
- Differential GPS (DGPS), 31
- documentation, 19, 34
 - database design, 38
 - validation procedures, 38
- domain schizophrenia, 41
- Domain value redundancy, 40
- duplicate data records, 41
- duplication
 - minimisation of, 18
- edit controls, 17
- education, 19
- error, 6
 - documentation of, 35
 - patterns, 40
 - visualisation, 46
- error prevention, 8, 10
- Federal Geographic Data Committee (FGDC), 26

feedback, 19
 Field data domain, 25
 fitness for use, 4, 34
 flexibility, 15
 gazetteers
 electronic, 33
 geodetic datums, 6, 42, 43
 Geodetic Networks, 26
 geo-referencing, 32, 42
 Georeferencing Guidelines, 33
 Geospatial Positioning Accuracy
 Standards (GPAS), 26
 Global Positioning System (GPS), 25, 30
 identification precision, 23
 inconsistency, 23
 inconsistent data values, 41
 incorrect data values, 40
 Indigenous rights, 48
 Information for Spatial Information in
 Europe), 25
Information Management Chain, 10, 18
 information quality contamination, 41
 Intellectual Property Rights, 47
 ISO 19115 for Geographic Information –
 Metadata, 25
 legal responsibilities, 47
 lineage, 36
 logical consistency, 36
 MaPSTeDI Guidelines, 33
 metadata, 34
 missing data values, 40
 moral responsibilities, 47
 nomenclatural data, 21
 nonatomic data values, 41
 outlier detection, 17
 partnerships, 13
 performance measures, 16
 positional accuracy, 25, 35
 precision, 3
 documentation of, 35
 false, 26
 numerical, 3
 statistical, 3
 primary species data, 3
 principles of data quality, 8
 prioritisation, 13
 privacy legislation, 47
 quality, 4
 quality assurance, 5
 quality control, 5
 Real-time Differential GPS, 31
 resolution, 3
 risk assessment, 47
 selective availability, 31
 spatial accuracy, 25
 spatial data, 25, 41
 Spatial Data Transfer Standards, 25
 species-occurrence data, 3
 Structure of Descriptive Data, 29
 targets
 setting of, 17
 Taxonomic Databases Working Group
 (TDWG), 29
 Taxonomic Impediment, 21
 taxonomy, 21
 temporal accuracy, 37
 threatened species, 48
 timeliness, 14
Total Data Quality Management cycle, 11
 trade sensitive species, 48
 training, 19
 transparency, 16
 truth in labelling, 7, 47
 uncertainty, 6
 update frequency, 14
 User Interface, 32
 validation, 6
 voucher collections, 24
 Wide Area Augmentation System
 (WAAS), 31

Chapter 4

Data Cleaning

Introduction to Chapter 4	1
Definition: Data Cleaning.....	1
The Need for Data Cleaning.....	2
Where are the Errors?.....	2
Preventing Errors.....	3
Spatial Error	3
Nomenclatural and Taxonomic Error	4
Merging Databases	4
Principles of Data Cleaning.....	5
Methods of Data Cleaning.....	8
Taxonomic and Nomenclatural Data.....	10
Identification certainty.....	10
Spelling of names	12
Spatial Data	25
Data Entry and Georeferencing	26
Geocode checking and validation.....	38
Descriptive Data.....	54
Documentation of Error.....	55
Visualisation of Error	56
Cited Tools	58
Software resources.....	58
On-line resources.....	60
Standards and Guidelines	61
Conclusion.....	63
References	64
Index to Chapter 4	70

This Chapter is equivalent to:

Chapman, A. 2005. *Principles and Methods of Data Cleaning*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 77 pp. ISBN: 87-92020-04-6 (available as a standalone PDF from <http://www.gbif.org>)

Introduction to Chapter 4

Data Cleaning is an essential part of the Information Management Chain as mentioned in the associated document, *Principles of Data Quality* (Chapman 2005a). As stressed there, error prevention is far superior to error detection and cleaning, as it is cheaper and more efficient to prevent errors than to try and find them and correct them later. No matter how efficient the process of data entry, errors will still occur and therefore data validation and correction cannot be ignored. Error detection, validation and cleaning do have key roles to play, especially with legacy data (e.g. museum and herbarium data collected over the last 300 years), and thus both error prevention and data cleaning should be incorporated in an organisation's data management policy.

One important product of data cleaning is the identification of the basic causes of the errors detected and using that information to improve the data entry process to prevent those errors from re-occurring.

This document will examine methods for preventing as well as detecting and cleaning errors in primary biological collections databases. It discusses guidelines, methodologies and tools that can assist museums and herbaria to follow best practice in digitising, documenting and validating information. But first, it will set out a set of simple principles that should be followed in any data cleaning exercises.

Definition: Data Cleaning

A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions.

The general framework for data cleaning (after Maletic & Marcus 2000) is:

- Define and determine error types;
- Search and identify error instances;
- Correct the errors;
- Document error instances and error types; and
- Modify data entry procedures to reduce future errors.

There are a number of terms used by different people to refer largely to the same process. It is a matter of preference what one uses. Terms include:

- Error Checking;
- Error Detection;
- Data Validation;
- Data Cleaning;
- Data Cleansing;
- Data Scrubbing; and
- Error Correction.

Introduction to Chapter 4

Data Cleaning is an essential part of the Information Management Chain as mentioned in the associated document, *Principles of Data Quality* (Chapman 2005a). As stressed there, error prevention is far superior to error detection and cleaning, as it is cheaper and more efficient to prevent errors than to try and find them and correct them later. No matter how efficient the process of data entry, errors will still occur and therefore data validation and correction cannot be ignored. Error detection, validation and cleaning do have key roles to play, especially with legacy data (e.g. museum and herbarium data collected over the last 300 years), and thus both error prevention and data cleaning should be incorporated in an organisation's data management policy.

One important product of data cleaning is the identification of the basic causes of the errors detected and using that information to improve the data entry process to prevent those errors from re-occurring.

This document will examine methods for preventing as well as detecting and cleaning errors in primary biological collections databases. It discusses guidelines, methodologies and tools that can assist museums and herbaria to follow best practice in digitising, documenting and validating information. But first, it will set out a set of simple principles that should be followed in any data cleaning exercises.

Definition: Data Cleaning

A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions.

The general framework for data cleaning (after Maletic & Marcus 2000) is:

- Define and determine error types;
- Search and identify error instances;
- Correct the errors;
- Document error instances and error types; and
- Modify data entry procedures to reduce future errors.

There are a number of terms used by different people to refer largely to the same process. It is a matter of preference what one uses. Terms include:

- Error Checking;
- Error Detection;
- Data Validation;
- Data Cleaning;
- Data Cleansing;
- Data Scrubbing; and
- Error Correction.

I tend to use the term *Data Cleaning* to encompass three sub-processes, viz.

- Data checking and error detection;
- Data validation; and
- Error correction.

A fourth – improvement of the error prevention processes – could perhaps be added.

The Need for Data Cleaning

The need for data cleaning is centred around improving the quality of data to make them “fit for use” by users through reducing errors in the data and improving their documentation and presentation (see associated document on *Principles of Data Quality* – Chapman 2005a). Errors in data are common and are to be expected. Redman (1996) suggested that unless extraordinary efforts have been taken, that a field error rate of 1-5% should be expected. The usual view of errors and uncertainties is that they are bad, but a good understanding of errors and error propagation can lead to active quality control and managed improvement in the overall data quality (Burrough and McDonnell 1998). Errors in spatial position (geocoding) and in identification are two of the major causes of error in species-occurrence data and it is the cleaning of these errors that is covered in this paper. Correcting errors in data and eliminating bad records can be a time consuming and tedious process (Williams *et al.* 2002) but it cannot be ignored. It is important, however, that errors not just be deleted, but corrections documented and changes traced. As mentioned in the companion document on *Principles of Data Quality*, it is best to add corrections to the database while retaining the original data in a separate field or fields so that there is always the chance of going back to the original information.

Where are the Errors?

Primary species data encompass a whole range of data – from museum and herbarium data, through observational data (point-based, regional or area-based, and systematic or grid-based), to survey data, both systematic and other (Chapman 2005a). Because of the historical nature of many museum and herbarium collections (often termed legacy data); many records carry little geographic information other than a general description of the location where they were collected (Chapman and Milne 1998). With historical data, where geocodes coordinates are given they are often not very accurate (Chapman 1999) and have generally been added at a later date by those other than the collector (Chapman 1992). Many of these data have drawbacks when it comes to use for species’ distribution studies. Observational and survey data are also valuable records for many studies and the georeferencing information may be quite accurate, but because vouchered reference material is seldom retained, the taxonomic or nomenclatural information are generally less reliable than for documented museum collections. The georeferencing of survey and observational records may, however, still include errors or ambiguities, for example it may not be clear as to whether the geocode refers to the centre of the grid, or one corner in grid-based records.

Much of the data (both museum and observational) have been collected opportunistically rather than systematically (Chapman 1999, Williams *et al.* 2002) and this can result in large spatial biases – for example, collections that are highly correlated with road or river networks (Margules and Redhead 1995, Chapman 1999, Peterson *et al.* 2002, Lampe and Riede 2002). Museum and herbarium data and most observational data, generally only supply information on the presence of the entity at a particular time and says nothing about absences in any other place or time (Peterson *et al.* 1998). This restricts their use in some environmental models,

but they remain the only complete collection of biological information covering the last 200+ years. The cost of replacing these data with new surveys would be prohibitive. It is not unusual for a single survey to exceed \$1 million (Burbidge 1991). Further, because of their collection over time, they provide irreplaceable baseline data about biological diversity during a time when humans have had tremendous impact on such diversity (Chapman and Busby 1994). They are an essential resource in any effort to conserve the environment, as they provide the only fully documented record of the occurrence of species in areas that may have undergone habitat change due to clearing for agriculture, urbanisation, climate change, or been modified in some other way (Chapman 1999).

Preventing Errors

As stressed previously, the prevention of errors is preferable to their later correction, and new tools are being developed to assist institutions in preventing errors.

Tools are being developed to assist the process of adding georeferencing information to databased collections. Such tools include eGaz (Shattuck 1997), geoLoc (CRIA 2004a), BioGeomancer (Peabody Museum *n.dat.*), GEOLocate (Rios and Bart *n.dat.*) and the Georeferencing Calculator (Wieczorek (2001b)). A project funded in 2005 by the Gordon and Betty Moore Foundation and involving worldwide collaboration is now bringing many of these tools together with the aim of making them available both as stand-alone open-source software tools and as Web Services. The need for further and more varied validation tools cannot, however, be denied. These tools will be discussed more fully later in this paper.

Tools are also being developed to assist in reducing error with taxonomic and nomenclatural data. There are two main causes of error with these data. They are inaccurate identifications or misidentifications (the taxonomy) and misspellings (the nomenclature). Tools to assist with the identification of taxa include improved taxonomies, floras and faunas (both regional and local), automated and computer-based keys to taxa, and digital imaging of type and other specimens. With the spelling of names, global, regional and taxonomic name-lists are being developed which allow for the development of authority files and database entry checklists that reduce error at data entry.

Perhaps the best way of preventing many errors is to properly design the database in the first instance. By implementing sound Relational Database philosophy and design any information that is frequently repeated such as species' names, localities and institutions, need only be entered once, and verified at the outset. Referential integrity then protects the accuracy of future entries.

Spatial Error

In determining the quality of species data from a spatial viewpoint, there are a number of issues that need to be examined. These include the identity of the specimen or observation – a misidentification may place the record outside the expected range for the taxon and thus appear to be a spatial error – errors in the geocoding (latitude and longitude), and spatial bias in the collection of the data. The use of spatial criteria may help in finding taxa that may be misidentified as well as misplaced geographically. The issue of spatial bias – very evident in herbarium and museum data (e.g. collections along roads) is more an issue for future collections, and future survey design rather than being related to the accuracy of individual plant or animal specimen records (Margules and Redhead 1995, Williams *et al.* 2002). Indeed – collection bias is more related to the totality of collections of an individual species, than it is to any one individual collection. In order to improve the overall spatial and taxonomic

coverage of biological collections within an area, and hence reduce spatial bias, existing historical collection data can be used to determine the most ecologically valuable locations for future surveys for example, by using environmental criteria (climate etc.) as well as geographic criteria (Neldner *et al.* 1995).

Nomenclatural and Taxonomic Error

Names form the major key for accessing information in primary species databases. If the name is wrong, then access to the information by users will be difficult, if not impossible. In spite of having rules for biological nomenclature for around 100 years, the nomenclatural and taxonomic information in a database (the *Classification Domain* of Dalcin 2004) is often the most difficult in which to detect and clean errors. It is also the area that causes the most angst and loss of confidence amongst users in primary species databases. This is often due to ignorance amongst users of the need for taxonomic changes and nomenclatural changes, but is also partly due to taxonomists not fully documenting and explaining these changes to users, complications in the relationship between names and taxa, and confusion with taxonomic concepts that are often not well covered in primary species databases (Berendsohn 1997).

The easier of these errors to clean is the nomenclatural data – the misspellings. Lists of names (and synonyms) are the key tools for helping with this task. Many lists already exist for regions and/or taxonomic groups, and these are gradually being integrated into global lists (Froese and Bisby 2002). There are still many regions of the world and taxonomic groups, however that do not have reliable lists.

Taxonomic error – the inaccurate identification or misidentification of the collection is the most difficult of errors to detect and clean. Museums and herbaria have traditionally had a *determinavit* system in operation whereby experts working in taxonomic groups examine the specimens from time to time and determine their circumscription or identification. This is a proven method, but one that is time-consuming, and largely haphazard. There is unlikely to be any way around this, however, as automated computer identification is unlikely to be an option in the near or even long-term future. There are, however, many tools available to help with this process. They comprise both the traditional taxonomic publications with which we are all familiar and newer electronic tools. Traditional tools include publications such as taxonomic revisions, national and regional floras and faunas, and illustrated checklists. Newer tools include automated and computer-generated keys to taxa; interactive electronic publications with illustrations, descriptions, keys, and illustrated glossaries; character-based databases; imaging tools; scientific image databases that include images of types; systematic images of collections; and easily accessible on-line images (both scientifically verified and others).

Merging Databases

The merging of two or more databases will both identify errors (where there are differences between the two databases) and create new errors (i.e. duplicate records). Duplicate records should be flagged on merging so that they can be identified and excluded from analysis in cases where duplicate records may bias an analysis, etc., but should generally not be deleted. While appearing to be duplicates, in many cases the records in the two databases may include some information that is unique to each, so just deleting one of the duplicates (known as ‘Merge and Purge’ (Maletic and Marcus 2000)) is not always a good option as it can lead to valuable data loss.

An additional issue that may arise with merging databases is the mixing of data that are based on different criteria such as different taxonomic concepts, different assumptions or units of measurements and different quality control mechanisms. Such merging should always document the source of the individual data units so that data cleaning processes can be carried out on data from the different sources in different ways. Without doing this, it may make the database more difficult to clean effectively and to effectively document any changes.

Principles of Data Cleaning

Many of the principles of data cleaning overlap with general data quality principles covered in the associated document on *Principles of Data Quality* (Chapman 2005a). Key principles include:

Planning is Essential (Developing a Vision, Policy and Strategy)

Good planning is an essential part of a good data management policy. The Information Management Chain (figure 1) (Chapman 2005a), includes Data Cleaning as a central portion that needs to be incorporated into the organisation's data quality vision and policy. A strategy to implement data cleaning and validation into the organisation's culture will improve the overall quality of the organisation's data and improve its reputation with users and suppliers alike.

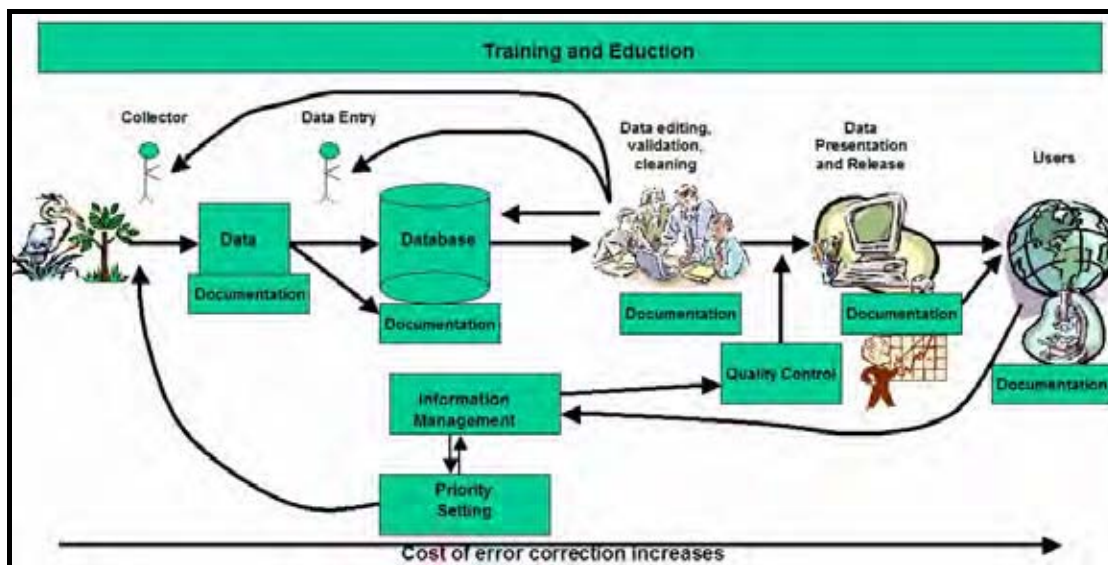


Fig. 1. *Information Management Chain showing that the cost of error correction increases as one moves along the chain. Education, Training and Documentation are integral to all steps (from Chapman 2005a).*

Organizing data improves efficiency

Organizing data prior to data checking, validation and correction can improve efficiency and considerably reduce the time and cost of data cleaning. For example, by sorting data on location, efficiency gains can be achieved through checking all records pertaining to one location at the same time, rather than going back and forth to key references. Similarly, by sorting records by collector and date, it is possible to detect errors where a record may be at an unlikely location for that collector on that day. Spelling errors in a variety of fields may also be found in this way.

Prevention is better than cure

As stressed previously (Chapman 2005a), it is far cheaper and more efficient to prevent an error, than to have to find it and correct it later. It is also important that when errors are detected, that feedback mechanisms ensure that the error doesn't occur again during data entry, or that there is a much lower likelihood of it re-occurring. Good database design will ensure that data entry is controlled so that entities such as taxon names, localities and persons are only entered once and verified at the time of entry. This can be done through use of drop-down menus or through keystroke identification of existing entries within a field.

Responsibility belongs to everyone (collector, custodian and user).

Responsibility for data cleaning belongs to all. The primary responsibility of the data-cleaning portion of the Information Management Chain (figure 1) obviously belongs to the data custodian – the person or organisation with principal responsibility for managing and storing the data. The collector, too, has responsibility and needs to respond to the custodian's questions when the custodian finds errors or ambiguities that may refer back to the original information supplied by the collector. These may relate to ambiguities on the label, errors in the date or location, etc. As will become obvious later in this document, the user also has a key responsibility to feed back to custodians information on any errors or omissions they may come across, including errors in the documentation associated with the data. It is often the user, when analysing or looking at the data in the context of other data, who will identify errors and outliers in the data that would otherwise go un-noticed. A single museum may have only a subset of the total available data (from one State or region for example), and it is only when the data are combined with data from other sources that errors may become obvious. Many of the tools elaborated in this document perform much better when looking at the totality of data for a species, collector or expedition than at subsets of them.

Partnerships improve Efficiency

Partnerships can be a very efficient method for managing data cleaning. As mentioned, the user is often the one who will be in the best position to identify errors in the data. If data custodians can develop partnerships with these key users then those errors won't be ignored. By developing partnerships, many data validation processes won't need to be duplicated, errors will more likely be documented and corrected, and new errors won't be incorporated by inadvertent "correction" of suspect records that are not in error. It is important to make these partnerships with users inside the organisation as well as outside as discussed in the associated paper on *Principles of Data Quality*.

Prioritisation reduces Duplication

As with organisation and sorting, prioritisation helps reduce costs and improves efficiency. It is often of value to concentrate on those records where extensive data can be cleaned at the lowest cost. For example, those that can be examined using batch processing or automated methods, before working on the more difficult records. By concentrating on those data that are of most value to users, there is also a greater likelihood of errors being detected and corrected. This improves client/supplier relationships and reputations, and provides greater incentive for both data suppliers and users to improve the quality of the data because it has an immediate use.

Setting of Targets and Performance Measures

Performance measures are a valuable addition to quality control procedures, and are used extensively with spatial metadata. They also help an organisation to manage their data cleaning processes. As well as providing users with information on the data and on their quality, such measures can be used by managers and curators to track those parts of the database that may need attention. Performance measures may include statistical checks on the data (for example, 95% of all records have an accuracy of less than 5,000 meters from their reported position), on the level of quality control (for example – 65% of all records have been checked by a qualified taxonomist within the previous 5 years; 90% have been checked by a qualified taxonomist within the previous 10 years), completeness (e.g. all 10-minute grid squares have been sampled) (Chapman 2005a).

Minimise duplication and reworking of data

Duplication is a major factor with data cleaning in most organisations. Many organisations carry out georeferencing at the same time as they database the record. As records are seldom sorted geographically, this means that the same or similar locations will be chased up a number of times. By carrying out the georeferencing of collections that only have textual location information and no coordinate information as a special operation, records from similar locations can then be sorted and located on the appropriate map-sheet or gazetteer. Some institutions also use the database itself to help reduce duplication by searching to see if the location has already been georeferenced (see under *Data Entry and Georeferencing*, below).

The documentation of validation procedures (preferably in a standardised format) is also important to reduce the reworking of data. For example, data quality checks carried out on data by a user may identify a number of suspect records. These records may then be checked and found to be valid records and genuine outliers. If this information is not documented in the record, further down the line, someone else may come along and carry out more data quality checks that again identify the same records as suspect. This person may then spend more valuable time rechecking the information and reworking the data. When designing databases, a field or fields should be included that indicates whether the data have been checked, by whom and when and with what result.

Experience in the business world has shown that the use of information chain management (see figure 1) can reduce duplication and re-working of data and lead to a reduction of error rates by up to 50% and a reduction in costs resulting from the use of poor data by up to two thirds (Redman 2001). This is largely due to efficiency gains through assigning clear responsibilities for data management and quality control, minimising bottlenecks and queue times, minimising duplication through different staff re-doing quality control checks, and improving the identification of better and improved methods of working (Chapman 2005a).

Feedback is a two-way street

Users of the data will inevitably carry out error detection, and it is important that they feedback the results to the custodians. As already mentioned, the user often has a far better chance of detecting certain error types through combining data from a range of sources, than does each individual data custodian working in isolation. It is essential that data custodians encourage feedback from users of their data, and implement the feedback that they receive (Chapman 2005a). Standard feedback mechanisms need to be developed, and procedures for receiving feedback agreed between the data custodians and the users. Data custodians also

need to convey information on errors to the collectors and data suppliers where relevant. In this way there is a much higher likelihood that the incidence of future errors will be reduced and the overall data quality improved.

Education and Training improves techniques

Poor training, especially at the data collection and data entry stages of the Information Quality Chain, is the cause of a large proportion of the errors in primary species data. Data collectors need to be educated about the requirements of the data custodian and users of the data so that the right data are collected (i.e. all relevant parts and life stages), that the collections are well documented – i.e. the locality information is well recorded (for example – does 10 km NW of Town ‘y’ mean 10 km by road, or in a direct line), that standards are applied where relevant (e.g. that the same grid size is used for related surveys), and that the labels are clear and legible and preferably laid out in a consistent manner to make it easier for data entry operators.

The training of data entry operators is also important as identified in the MaPSTeDI georeferencing guidelines (University of Colorado Regents 2003a). Good training of data entry operators can reduce the error associated with data entry considerably, reduce data entry costs and improve overall data quality.

Accountability, Transparency and Audit-ability

Accountability, transparency and audit-ability are essential elements of data cleaning. Haphazard and unplanned data cleaning exercises are very inefficient and generally unproductive. Within data quality policies and strategies – clear lines of accountability for data cleaning need to be established. To improve the “fitness for use” of the data and thus their quality, data cleaning processes need to be transparent and well documented with a good audit trail to reduce duplication and to ensure that once corrected, errors never re-occur.

Documentation

Documentation is the key to good data quality. Without good documentation, it is difficult for users to determine the fitness for use of the data and difficult for custodians to know what and by whom data quality checks have been carried out. Documentation is generally of two types and provision for them should be built into the database design. The first is tied to each record and records what data checks have been done and what changes have been made and by whom. The second is the metadata that records information at the dataset level. Both are important, and without them, good data quality is compromised.

Methods of Data Cleaning

Introduction

Museums and herbaria throughout the world are beginning to database their collections at increasing rates, and are starting to make at least some of that information available via the Internet. The rate of databasing of collections has increased in recent years with the development of tools and methodologies that can assist in the process, and increased publication since the creation of the Global Biodiversity Information Facility (GBIF) with its aim to “make the world's primary data on biodiversity freely and universally available via the Internet” (GBIF 2003a).

As well as good practices (see associated document – *Principles of Data Quality* and *Principles of Data Cleaning*, this document), there is a need for useful and powerful tools that automate, or greatly assist in the data cleaning process. Automated methods can only be part of the procedure and there is a continuing need for the development of new tools to assist this process, and for their use to be integrated into best practice routines. Manual cleaning of data is laborious and time consuming, and is in itself prone to errors (Maletic and Marcus 2000), but it will continue to be important with primary species-occurrence databases. Where possible, it should only be carried out as a last resort, for small data sets, or where other checks have left just a few errors that cannot be checked any other way.

Some of the techniques that have been developed include the use of climate models to identify outliers in climate space (Chapman 1992, 1999, Chapman and Busby 1994), in geographic space (CRIA 2004b, Hijmans *et al.* 2005, Marino *et al.* in prep.), the use of automated georeferencing tools (Beaman 2002, Wieczorek and Beaman 2002) and many others. Most collection institutions do not have a high level of expertise in data management techniques or in Geographic Information Systems (GIS). What is needed in these institutions is a simple, inexpensive set of tools to both assist in the input of data and information, including geocoding information, and similar simple and inexpensive tools for data validation that can be used without the necessary incorporation of expensive GIS software. Some tools have been developed to assist with data entry – tools such as Biota (Colwell 2002), BRAHMS (University of Oxford 2004), Specify (University of Kansas 2003a), BioLink (Shattuck and Fitzsimmons 2000), Biótica (Conabio 2002), and others that provide database management and associated data entry (Podolsky 1996, Berendsohn *et al.* 2003); and eGaz (Shattuck 1997), geoLoc (CRIA 2004a, Marino *et al.* in prep.), GEOLocate (Rios and Bart *n.dat.*) and BioGeoMancer (Peabody Museum *n.dat.*), that assist in the georeferencing of collections. There are also a number of documented guidelines available on the Internet that can assist institutions in setting up and managing their databasing programs. Examples include the MaNIS Georeferencing Guidelines (Wieczorek 2001a), MaPSTeDI Georeferencing Guidelines (University of Colorado Regents 2003a) and HISPID (Conn 1996, 2000).

There are many methods and techniques that can aid in the cleaning of errors in primary species and species-occurrence databases. They range from methods that have been operating in museums and herbaria for hundreds of years, to automated methods that are still largely untested. This paper looks in detail at a range of methods for cleaning species databases, and where possible, provides examples. It is by no means a comprehensive list as many institutions have developed their own techniques and methodologies.

Because of the very nature of natural history collections, it is not possible that all geocode information be highly precise, or that there is a consistent level of precision within a database. Data with a very low precision, however, are not necessarily of low quality. Quality only comes into being once the data are being used and is not a character of the data *per se* (see discussion in associated paper on *Principles of Data Quality* - Chapman 2005a). Quality is merely a factor of fitness for use or potential use and is a relative term. What is important is for users of the data to be able to determine from the data themselves, if the data are likely to be fit for the required application. The level of accuracy of each given geocode should therefore be recorded within the database. I prefer this to be in non-categorical form, recorded in meters, however many databases have developed categorical codes for this purpose. When this information is available, a user can request, for example, only those data that are better than a certain metric value – e.g. better than 5,000 meters (see example using codes for

extracting data at University of Colorado Regents 2003b). There are a number of ways of determining accuracy of geocoded records. The point-radius method (Wieczorek *et al.* 2004) is, I believe, the easiest and most practical method, and is one previously recommended for use in Australia (Chapman and Busby 1994). It is also important that automated georeferencing tools include calculated accuracy as a field in the output. The geoLoc (CRIA 2004a, Marino *et al.* in prep.) and BioGeomancer (Peabody Museum *n.dat.*) tools, which are still under development, include this feature.

Over time, it is hoped that species collection data resources will improve as institutions move to more precise instrumentation (such as GPS) for recording the location of new records and as historic records are corrected and improved. It is also important that collectors make the best possible use of the tools available to them and not just use a GPS to record data to 1 arc minute resolution because of historical reasons - as that is the finest they have recorded information prior to using a GPS. If this is done, then they should make sure that the appropriate accuracy is added to the database otherwise it may be assumed that as a GPS was used, the accuracy is 10 meters as opposed to the 2000 meters of reality. Error prevention is preferable to error detection, but the importance of error detection cannot be under stressed, as error prevention alone can never be guaranteed to prevent all possible errors.

Taxonomic and Nomenclatural Data

Names, whether they are scientific binomials or common names, provide the first point of entry to most species and species-occurrence databases. Errors in names may arise in a number of ways: the identification may be wrong, the name may be misspelt, or the format may be wrong (or not what is expected by the user). The first of these is not easy to check or rectify without tedious effort, and requires the services of a taxonomic expert. The others though, are more easily catered for with good database design and methods that assist with data entry so that these errors do not occur or are minimised.

Identification certainty

Traditionally, museums and herbaria have had an identification or “*determinavit*” system in operation whereby experts working in taxonomic groups from time to time examine the specimens and determine their identifications. This may be done as part of a larger revisionary study, or by an expert visiting another institution and, while there, checks the collections. This is a proven method, but one that is time-consuming, and largely haphazard. There is unlikely to be any way around this, however, as automated computer identification is unlikely to be an option in the near or even long-term future.

Database design

One option may be the incorporation of a field in databases that provides some indication of the certainty of the identification when made. There are a number of ways that this could be done, and it perhaps needs some discussion to develop a standard methodology. This would be a code field, and may be along the lines of:

- identified by World expert in the taxa with high certainty
- identified by World expert in the taxa with reasonable certainty
- identified by World expert in the taxa with some doubts
- identified by regional expert in the taxa with high certainty
- identified by regional expert in the taxa with reasonable certainty
- identified by regional expert in the taxa with some doubts

- identified by non-expert in the taxa with high certainty
- identified by non-expert in the taxa with reasonable certainty
- identified by non-expert in the taxa with some doubt
- identified by collector with high certainty
- identified by collector with reasonable certainty
- identified by collector with some doubt

How one might rank these would be open to some discussion, and likewise whether these were the best categories or not. There are apparently some institutions that already do have a field of this nature. The HISPID Standard Version 4 (Conn 2000) does include a simplified version – the Verification Level Flag with five codes (Table 1).

Many institutions also already have a form of certainty recording with the use of terms such as: “aff.”, “cf.”, “s. *lat.*”, “s. *str.*”, “?”. Although some of these (aff., cf.) have strict definitions, their use by individuals can vary considerably. The use of *sensu stricto* and *sensu lato* imply variations in the taxonomic concept rather than levels of certainty, although not always used in that way.

0	The name of the record has not been checked by any authority
1	The name of the record determined by comparison with other named plants
2	The name of the record determined by a taxonomist or by other competent persons using herbarium and/or library and/or documented living material
3	The name of the plant determined by taxonomist engaged in systematic revision of the group
4	The record is part of type gathering or propagated from type material by asexual methods

Table 1. Verification Level Flag in HISPID (Conn 2000).

As an alternative, where names are derived from other than taxonomic expertise, one could list the source of the names used (after Wiley 1981):

- descriptions of new taxa
- taxonomic revisions
- classifications
- taxonomic keys
- faunistic or floristic studies
- atlases
- catalogues
- checklists
- handbooks
- taxonomic scholarship/rules of nomenclature
- phylogenetic analysis

Data entry

As data are being entered into the database, checks can be made as to whether the name has been checked by an expert or not, and if any of the above fields are present, that they have been entered. If such fields are used, they should be entered through use of a check-list or

authority file that restricts the available options and thus reduces the chance of errors being added.

Error checking

Geocode checking methods (see under *Spatial Data*, below) can also help identify misidentifications or inaccurate identifications through the detection of outliers in geographic or environmental space. Although generally an outlier found through geocode checking will be an error in either the latitude or longitude, occasionally it indicates that the specimen has been misidentified as the taxon being studied and thus falls outside the normal climate, environmental or geographic range of the taxon. See below for a more detailed discussion on techniques for identifying geographic outliers.

The main method for detecting whether a collection is accurately identified or not, though, is for experts to check the identification by examining the specimen or voucher collections where they exist. Geocode outlier detection methods cannot determine if a collection is accurately identified or not, but may help identify priority collections for expert taxonomic checking. With observational data, experts may be able to determine, on personal knowledge, if the taxon is a likely record for the area (e.g. *Birds Australia 2004*); but generally it is difficult to identify an inaccurate identification of an observational record where there are no voucher specimens. Many institutions may flag doubtful or suspect records and then it is up to the user to decide if they are suitable for their use or not.

Spelling of names

This paper does not attempt to cover all the possible kinds of names that may be entered into a primary species database. For example, hybrids and cultivars in plant databases, synonyms of various kinds, and taxonomic concepts all have specific issues and the checking of these can be problematic. Examples of how such names may be treated can be found in the various International Codes of Nomenclature, as well as in TDWG Standards such as HISPID (Conn 1996, 2000) and Plant Names in Botanical Databases (Bisby 1994).

Scientific names

The correct spelling of a scientific name is generally governed by one of the various relevant nomenclature Codes. However, errors can still occur through typing errors, ambiguities in the Code, etc. The easiest method to ensure such errors are kept to a minimum is to use an 'Authority File' during input of data. Most databases can be set up to incorporate either an unchangeable authority file, or an authority file that can be updated during input.

Database design

One of the keys to being able to maintain good data quality with taxon names is to split the name into individual fields (i.e. atomise the data) rather than maintain them all in one field (e.g. genus, species, infraspecific rank, infraspecific name, author and certainty). Maintaining these in one field reduces the opportunities for good data validation and error detection, and can lead to a quite considerable increase in the opportunity for errors. For example, by separating the genus and species parts of the name, each genus name only needs to be added once into a relational database (through an authority file or pick-list), thus reducing opportunities for typographic errors and misspellings.

Databases that include all parts of the name in one field can make it very difficult to maintain quality or to combine with other databases. It introduces another range of possible errors and

is not recommended. Some databases use both – a combined field and atomised fields, but this again provides opportunity for added error if these are not automatically generated, and if one is updated and the other not. Automatically generated fields eliminate the danger of this.

Two issues that need to be considered with atomised data are the incorporation of data in a database where the data is in one field (for example the importing of lists of plants or animals), and the need to present a concatenated view of data from an atomised database, for example on a web site or in a publication.

With the first of these – the parsing of data from a concatenated field into individual fields is generally not as simple a process as it might appear. It is not only an issue with names, of course, as the same problems arise with references and locality information as discussed below. As far as I am aware, no simple tools for doing the parsing exist, however many museums and herbaria have done this for their own institutions, and thus algorithms may be available from these institutions.

With the second, the requirement to present a concatenated view on the output side for presentation on the Web or in reports could either be carried out using an additional (generated) field within the database that concatenates the various fields, or done on the fly during extraction. This is an issue that should be considered when designing the database and its reporting mechanisms. They are issues that the taxonomic community may need to discuss further with the aim of developing simple tools or methodologies.

Authority files

Authority files exist for a number of taxonomic groups, and are being developed by a range of agencies. Reliable authority files are available for many higher taxa (Families, Orders, and Genera), and these can be used to ensure data integrity in these fields. It is unlikely that a detailed authority file for all taxa, especially to the species level and below, will be produced in the near future, however, existing authority files (e.g. IPNI 1999, Froese and Bisby 2004) can be used as a beginning. If authority files are available, then the databases can be set up in such a way that new names can be added to them. For example, assume a database has an authority file with a pull down list, or fills in the field as one types (for example as happens in an EXCEL spreadsheet if one starts to type a name in a field where that name may already be in an earlier row).

1. Use the pull down list to search for the name
2. It is not there
3. Click on the button – “New name”
4. Add the New Name
5. The database may come back and say “This name is similar to <name>” do you want to continue?
6. Yes
7. The name is added to the list, and the next time you wish to add a name, that name will now appear in the pull-down list.

In this way, you are gradually adding to and improving the authority file.

As an extra check, these names may then go into a secondary list that a supervisor can verify and either approve or discard. Depending on the level of sophistication of the database, the list may include synonyms and if you begin to type in a name, it may ask you if you really wish to add this name as it is listed in the authority file as a synonym of <name>.

It is recommended that Authority files be used wherever possible. A good start is the Species2000 & ITIS Catalogue of Life (Froese and Bisby 2004), available on CD as an annual checklist. The format of this document is being improved for future editions to make it easier to incorporate into databases. The checklist is also available electronically for checking individual taxa and is in addition to a regularly updated checklist, which is also available, on-line. Also, a number of names databases exist or are being developed and these can form the basis of a names authority file. Some examples include,

Global lists such as:

- Species2000 & ITIS Catalogue of Life (Froese and Bisby 2002),
- Ecat (GBIF 2003b),
- International Plant Name Index (IPNI 1999);
- Global Plant Checklist (IOPI 2003).

Regional lists such as:

- Integrated Taxonomic Information System (Ruggiero 2001);
- Australian Plant Name Index (Chapman 1991, ANBG 2003);
- Proyecto Anthos – Sistema de información sobre los plantas de España (Fundación Biodiversidad 2005)
- Australian Faunal Directory (ABRS 2004);
- Med Checklist (Greuter *et al.* 1984-1989).

Taxonomic lists such as:

- ILDIS World Database of Legumes (Bisby *et al.* 2002);
- Fishbase (Froese and Pauly 2004);
- World Spider Catalog (Platnik 2004);
- Many others.

Where authority files are imported from an external source such as one of those above, then the Source-Id should be recorded in the database so that changes that are made between editions of the authority source can be easily incorporated into the database, and the database updated. Hopefully, before long this may become easier through the use of Globally Unique Identifiers (GUIDs)¹.

Duplicate entries

Even when designing a database from scratch and trying to normalise it as much as possible for example by using authority tables, the issue of duplicate records cannot be avoided, and especially when importing data from secondary sources (e.g. names or references). To remove (or flag) such duplicates a special interface may be needed. The interface should be capable of identifying potential duplicates using special algorithms. The data entry operator (or curator, expert, etc.) will then have to decide from the list of potential duplicates the set of records identified as real duplicates and the records that should be retained. The systems will then discard and archive, or flag the superfluous records while keeping the referential integrity of the system. Generic software could be implemented to rectify this, but as with the parsing software, does not appear to exist at the moment. Biodiversity database designers should be aware of the problem and consider designing a generic software tools for these tasks.

¹ <http://www.webopedia.com/TERM/G/GUID.html>

Error checking

It is possible to carry out some automated checks on names. Although complete lists of species names do not exist, lists of family names, and generic names (e.g. IAPT 1997, Farr and Zijlstra *n.dat.*) are much more complete, especially for some taxa. Checks against these fields could be carried out against such lists. With species epithets (the second part of the binomial) there are a number of tests that can be conducted. For example, looking for names within the same genus that have a high degree of similarity – names with one character out of place or with a character or characters missing, etc. The CRIA Data Cleaning system (CRIA 2005) carries out many of these tests on distributed data obtained through speciesLink (CRIA 2002). Best practice in this case would be for automated detection, but not automated correction. Other possible checks include (modified from English 1999, Dalcin 2004):

Missing Data Values – This involves searching for empty fields where values should occur. For example, in a botanical database if an infraspecific name is cited, then a value should be present in the corresponding infraspecific rank field; or if a species name is present, then a corresponding generic name should also be present.

Incorrect Data Values – This involves searching for typographic errors, transposition of key strokes, data entered in the wrong place (e.g. a species epithet in the generic name field), and data values forced into a field that requires a value (i.e. is a mandatory field), but for which the data entry operator doesn't know the value so adds a dummy value. There are a number of ways of checking for some of these errors – for example, using Soundex, (Roughton and Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), or Skeleton-Key (Pollock and Zamora 1984). Each of these methods uses slightly different algorithms for detecting similarity. A recent test of a number of methods (including those mentioned) using species names and a number of datasets (Dalcin 2004) showed that the Skeleton-Key method produced the highest proportion of true errors to false errors in the datasets tested. An on-line example of using these methods can be seen on the CRIA site in Brazil (CRIA 2005). These are further explained below.

Nonatomic Data Values – This involves searching for fields where more than one fact is entered. For example “subsp. bicostata” in the infraspecies field where this should be split into two fields. Depending on database design (see above) this may not be an error. Nonatomic data values occur in many databases and are difficult to remove. An essential first step is that such values indicate that the database probably needs a new field created. Some nonatomic data can then be split into the relevant fields using automated methods, but more often than not, many are left that can only be fixed manually under the control of an expert.

Domain Schizophrenia – This involves searching for fields used for purposes for which they may not have been intended. This often happens where a certainty field has not been included in the database and question marks, uncertainties such as cf., aff. are added in the same field as the species epithet, or comments added (Table 2). The nature of this ‘error’ may also depend on database design.

Family	Genus	Species
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

Table 2. Examples of Domain schizophrenia (from Chapman 2005a).

Duplicate Occurrences – This involves searching for names that may refer to the same real world value. There are two main types of duplicates that can occur here – the first is an error due to misspellings, and the second is where there is more than one valid alternate name such as with the International Code of Botanical Nomenclature (2000) which allows for alternate Family names (e.g. Brassicaceae/Cruciferae, Lamiaceae/Labiatae). The latter can be handled by either choosing one of the valid alternatives for the database, or using linked synonyms depending on the policy of the institution. Similar issues may also occur where alternate classifications have been followed at higher taxonomic ranks, or even at the genus level where a species may validly occur in more than one genus depending on whose classification is followed (*Eucalyptus/Corymbia*; Albatross species in the Southern Hemisphere, small wild cat species, and many more).

Inconsistent Data Values – This occurs where two related databases do not use the same names lists, and when combined (or compared) show inconsistencies. For example, this may occur at botanic gardens between the Living Collection and the Herbarium; when merging databases of two specialists; or in museums between the collection database and the images database. Correcting involves checking one database against the other to identify the inconsistencies.

Dalcin (2004) conducted a number of detailed experiments on methods of checking for spelling errors in scientific names and developed a set of tools to check for phonetic similarity. I have not used or tested these tools, but details on the methods and the results and comparative tests between methods can be obtained from Dalcin (2004) pp. 92-148. Also, CRIA, in Brazil have developed name-checking routines along similar lines (CRIA 2005) and these are expanded in the methods section below.

Common names

There are no hard and fast rules for ‘common’ or vernacular names, be they in Portuguese, Spanish, English, Hindi, various other languages, or regionally-based indigenous names. Often what are called ‘common’ names are in reality colloquial names (especially in botany) and may have just been coined from a translation of the Latin scientific name. In some groups, for example birds (see Christidis and Boles 1994) and fish (Froese and Pauly 2004), agreed conventions and recommended English names have been developed. In many groups the same taxon may have many common names, which are often region-, language-, or people-specific. An example is the species *Echium plantagineum* which is known variously as ‘Paterson’s Curse’ in one Australia State and ‘Salvation Jane’ in another and with other names (e.g. Viper’s Bugloss, Salvation Echium) in other languages and countries. Conversely, the same common name may be applied to multiple taxa, sometimes in different regions, but sometimes even in the same region.

It is just about impossible to standardise common names, even across one language except perhaps for some small groups. But does it make any sense to try and do this (Weber 1995)? True common names are names that have developed and evolved over time, and the purpose of having them is so people can communicate. What I am suggesting here is that common names not be standardised, but that when placed in a database it is done in a standard way and their source documented. Many users of primary species occurrence data want to access data through the use of common names, so there is value in having them in our databases if we want to make our data of the most use to the largest possible audience. By adopting standard methods for recoding common names, be it one per species or hundreds - in one

language or in many – and documenting the source of each name, we can make searching and thus information retrieval that much more efficient and useful.

There are many difficulties in including common names in species databases. These include:

- names in non-Latin languages that require the use of Unicode within the database for storage. Problems may occur:
 - where databases attempt to store the names phonetically using just the Latin alphabet,
 - where people are not able to display the characters properly on their screen or in print,
 - in carrying out searches where users have only a “Latin” keyboard,
 - with data entry where names are a mix of Latin and non-Latin,
- the need to store information on the language of the name, especially where names of mixed language are included,
- the need to store information on regional factors – the area for which the name may be relevant, the language dialect, etc.
- the need to store information such as the references to the source of the name.

It is not any easy task to do properly and particularly in a way that increases usefulness while reducing error. If it is decided to include such names, the following may help in providing some degree of standardisation.

Data entry

When databasing common names, it is recommended that some form of consistency in construction be followed. It is probably most important that each individual database be internally consistent. The development of regional or national standards is recommended where possible. There are too many languages and regional variations to attempt to develop a standard for all languages and all taxa, although some of the concepts proposed here could form the basis for a standard in languages other than those covered.

For English and Spanish common names, I recommend that a similar convention to that developed for use in Environment Australia (Chapman *et al.* 2002, Chapman 2004) be followed. These guidelines were developed to support consistency throughout the organisation’s many databases. These conventions include beginning each word in the name with an initial capital.

Sunset Frog

With generalised or grouped names a hyphen is recommended. The word following the hyphen is generally not capitalised, except for birds where the word following the hyphen is capitalised if it is a member of a larger group as recommended by Christidis and Boles (1994).

Yellow Spider-orchid

Double-eyed Fig-Parrot (‘Parrot’ has an initial capital as it is a member of the Parrot group).

Portuguese common names are generally given all in lower case, usually with hyphens between all words if a noun, or separated by a space if a noun and adjective. It is recommend that for Portuguese common names, either this convention be followed or be modified to conform to the English and Spanish examples.

mama-de-cadela,
fruta-de-cera
cedro vermelho

There is some disagreement and confusion as to whether apostrophes should be used with common names. For geographic names, there is a growing tendency to ignore all apostrophes (e.g. Smiths Creek, Lake Oneill), and it is now accepted practice in Australia (ICSM 2001, Geographic Names Board 2003 Art. 6.1). I recommend that a similar convention could be adopted with common names, although there is no requirement at present to do so.

Where names are added in more than one language and/or vary between regions, dialects or native peoples, then the language and the regional information should be included in a way that it can be attached to the name. This is best done in a relational database by linking to the additional regional and language fields, etc. In some databases, where there is only a language difference, the language is often appended to the name in brackets, but although this may appear to be a simple solution in the beginning, it usually becomes more complicated over time and often becomes unworkable. It is best to design the database to cater for these issues in the beginning rather than have to add flexibility at a later date.

If names in non-Latin alphabets are to be added to the database, then the database should be designed to allow for the inclusion of the UNICODE character sets.

Error checking

As common names are generally tied to the scientific name, checks can be carried out from time to time to check for consistency within the database. This can be a tedious procedure, but only need be carried out at irregular intervals. Checks can be done by extracting all unique occurrences and checking for inconsistencies, e.g. missing hyphens.

Again, programs such as Soundex, (Roughton and Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), or Skeleton-Key (Pollock and Zamora 1984) could be used to search for typographic errors such as transposition of characters as mentioned above for *scientific names* (see Dalcin 2004).

Infraspecific rank

The use of an infraspecific rank field(s) is a more significant in databases of plants than in databases of animals. Animal taxonomists general only use one rank below species – that of subspecies (and even this is used with decreasing frequency), with the name treated as a trinomial:

Stipiturus malachurus parimeda.

Historically, however, some animal taxonomists did use ranks other than subspecies, and databases may need to cater for these. If so, then the comments made for plant databases below, will also apply to databases of animal names.

Database design

As mentioned elsewhere, there are major data quality advantages in keeping the infraspecific rank separate from the infraspecific name. This allows for simple checking of the infraspecific rank field, and makes checking of the infraspecific name field easier as well. The rank and the name should never be included as “content” within the one field.

One issue that should be considered with atomised databases is the need in some cases to concatenate these fields for display on the web, etc. This can generally be automated, but consideration to how this may be done (whether in the database as an additional generated field or on the fly) should be considered when designing the database and its output forms.

Data entry

With plants (and historically animals), there are several levels below species that may be used. These infraspecific ranks are most frequently *subspecies*, *variety*, *subvariety*, *forma* and *subforma* (the Botanical Code does not preclude inserting additional ranks, so it is possible that other ranks may exist in datasets). Subvariety and subforma are seldom used, but do need to be catered for in plant databases. Again, a pick-list should be set up with a limited number of choices. If this is not done, then errors begin to creep in, and you will invariably see subspecies given as: subspecies, subsp., ssp., subspp., etc. This can then be a nightmare for anyone trying to extract data, or to carry out error checking. It is better to restrict the options at the time of input, than have to cater for a full range at the time of data extraction, or attempt data cleaning to enforce consistency at a later date. It is recommended that the following be used:

subsp.	subspecies
var.	variety
subvar.	subvariety
f.	form/forma
subf.	subform
cv.	cultivar (often treated in databases as another rank, but see separate comments below)

In collection databases, the inclusion of a hierarchy is not necessary where more than one level may exist, because this just adds an extra layer of confusion and under the International Code for Botanical Nomenclature (2000) the hierarchy is unnecessary to unambiguously define the taxon. If the hierarchy is included, it must be possible to extract only that which is necessary to unambiguously define the taxon.

Leucochrysum albicans subsp. *albicans* var. *tricolor* (= *Leucochrysum albicans* var. *tricolor*).

Error checking

If the database has been designed well and a checklist of values used, then there is less need for further error checking. Where this is not the case, however, checks should be carried out to ensure that only the limited subset of allowed values occurs. One check that should be done however is for missing values.

Cultivars and Hybrids

Cultivars and hybrids occur in many plant databases and are often not handled well. Cultivars are subject to their own Code of Nomenclature (Brickell *et al.* 2004). In many plant species databases they are treated as just another infraspecific rank (“cv.”) and in some databases this may be quite acceptable. Hybrids are much more difficult to handle than most other groups. They may be given a binomial name and can then be treated as any other taxon of the same rank (preceded by an “X” (multiplication sign) to denote a hybrid), or they may be treated as

a formula (a cross between two, or more taxa which may even be at different ranks) indicated with taxonomic names separated by multiplication signs.

Database design

I recommend that anyone looking at setting up a database of plants that may include hybrids or cultivars consult the HISPID standard (Conn 1996, 2000) where hybrids are treated as part of the Record Identification Group. However they are handled, it is good practice to include a field that states that the name belongs to a cultivar or hybrid, etc. In this way they can be extracted separately and treated differently (for formatting, concatenation, etc.) on extraction, and for error checking.

Error checking

Checking of errors for hybrids and cultivars is a difficult task if the database has not been set up to cater for it. One suggestion for checking may be to treat them as a group, i.e. extract all hybrids and sort them alphabetically by species depending on how they are stored in the database. This is much easier to do where a separate field is included that identifies hybrid records as such. One key error that is likely to occur is inconsistency with the use of the ‘X’ sign. Some databases may not allow for a multiplication sign and it is commonly replaced by an ‘x’ or ‘X’ sometimes with a space before the name and sometimes not. These sorts of inconsistencies can easily be checked. I know of no really good system for checking errors in hybrid names.

Unpublished Names

Data entry

Not all records placed in a primary species databases are going to belong to a validly published taxon. To be able to retrieve these records from the database it is necessary to provide a ‘temporary’ name for that collection. If unpublished names can be incorporated into a database in a standard format, it makes it a lot easier to keep track of them, and to be able to retrieve them at a later date. It is also better, and less confusing than adding unpublished names that are binomials that look like published names, with or without the tag such as “nomen novum”, “nom. nov.” and “ms”. Too often the “ms” or “nom. nov.” is left off and users can spend a lot of time looking for the publication and reference information for the unpublished name. By using a formula it is obvious to all that it is an unpublished name.

In the 1980s in Australia, botanists agreed on a formula (Croft 1989, Conn 1996, 2000) for use with unpublished names. This was to avoid confusion arising through the use of such things as “*Verticordia* sp.1”, “*Verticordia* sp.2” etc. Once databases begin to be combined, for example through the Australian Virtual Herbarium (CHAH 2002), *speciesLink* (CRIA 2002), Biological Collection Access Service for Europe (BioCASE 2003), the Mammal Networked Information System (MaNIS 2001), the GBIF Portal (GBIF 2004) and many others, names like these can cause even more confusion as there is no guarantee that what was called “sp.1” in one institution is identical to “sp.1” in a second. One way to keep these databases clean and consistent, and enable the smooth transfer of data from one to another, is through the use of a formula similar to that adopted by Australian botanical community.

The agreed formula is in the form of:

“<Genus> sp. <colloquial name or description> (<Voucher>)”:

Prostanthera sp. Somersbey (B.J.Conn 4024)

Later, when the taxon is formally described and named, the formula-name can be treated as a synonym, just like any other synonym.

The use of such a formula makes a database more complicated than it may otherwise be, because instead of the species field only ever having one word; to cater for the formula it now requires inclusion of a sentence. The use of “sp. 1” “nom. nov.” etc. as is often used have the same problem, and this method leaves less room for ambiguity. The use of formulae like these can cause difficulties with concatenation (for presentation on the web, etc.), however experience with the use of this methodology in Australia (for example, see the use with the SPRAT database of the Australian Department of the Environment (DEH 2005b)) has proved to work well. In all other ways, however, the formula is treated as a ‘species’ epithet, albeit with spaces and brackets, etc.

Because of the need to use unpublished names, for example in legal lists of threatened species (see for example, DEH 2005a), it is essential that there is a consistent system of naming or tagging these taxa for use in non-taxonomic publications, for example in legislative instruments. By using a formula like that suggested here, there is little danger of accidentally publishing a nomen nudum by mistake.

It is recommended that museum and herbaria adopt a similar system for use in their databases.

Error checking

The most common error that occurs with a formula name such as suggested here is that of misspelling. Because the formula usually includes several words, it is often easy to make a mistake with citation of the voucher, etc. The easiest way in which to check such names is to sort each within a genus (they should be the only names in the species or infraspecies fields with more than one word) and examine them for similarities. This should not be too onerous a task as there is unlikely to be a huge number within any one genus. Similar techniques such as Soundex, mentioned above, could also be used.

Author names

The authors of species names may be included in some specimen databases, but more often than not, their inclusion can lead to error as they are seldom thoroughly checked before inclusion. They are only really necessary where the same name may have inadvertently have been given to two different taxa (homonyms) within the same genus or where the database attempts to include taxonomic concepts (Berendsohn 1997). The inclusion of the author’s name following the species (or infraspecies) name can then distinguish between the two names or concepts. If databases do include authors of species names, then these should definitely be included in fields separate from the species’ names themselves.

The concatenation of data where author names and dates are kept separate is usually not a major issue except in plants with autonyms (see below). Mixed databases of plants and animals, however may cause some problems where authorities are treated slightly differently. It should not present too many difficulties if the author fields are set up in the database to cater for these but rules for extraction may need to be different for the different Kingdoms.

Dalcin (2004) treats the authority as a label element under his nomenclatural data domain.

Data entry

With animal names the author name (usually in full) is always followed by a year; with plants, the author name or abbreviation is given alone.

Animals:

Emydura signata Ahl, 1932

Macrotis lagotis (Reid, 1937)

(the bracket indicates that Reid ascribed the species to a different genus)

Plants:

Melaleuca nervosa (Lindley) Cheel

synonym: *Callistemon nervosus* Lindley

(Lindley originally described it as a *Callistemon*; Cheel later transferred it to the genus *Melaleuca*).

With plants, occasionally the terms “ex” or “in” may be found in author names. The author in front of the “ex” - the pre-'ex' author is one who supplied the name but did not fulfil the requirements for valid publication or who published the name before the nomenclatural starting date for the group concerned. A post-'in' author is one in whose work a description or diagnosis supplied by another author is published. For a further explanation of pre-'ex' and post-'in' authors and their use see Arts 46.2 and 46.3 of the International Code of Botanical Nomenclature (2000). If author names are used within databases they should be in separate fields to the name (see discussion on atomisation, above) and it is recommended that neither the pre-'ex' nor the post-'in' authors be cited.

Green (1985) ascribed the new combination *Tersonia cyathiflora* to "(Fenzl) A.S. George"; since Green nowhere mentioned that George had contributed in any way, the combining author must be cited as “A.S.George ex J.W.Green” or preferably as just “J.W.Green”.

Tersonia cyathiflora (Fenzl) J.W.Green

In W.T.Aiton's 2nd edition of *Hortus Kewensis* (1813), many of the descriptions are signed Robert Brown, and thus it can be assumed that Brown described the species. The author of the names is often cited as “R.Br. in Ait.” It is recommended, however that the author be cited as just “R.Br.”

Acacia acicularis R.Br.

With plants – for the type subspecies or variety, etc. where the infraspecific name is the same as the species name (autonym), the author of the species name is used and follows the specific epithet. This format regularly causes confusion for reconstruction of names in specimen databases that include author names, as it is an exception to other rules.

Leucochrysum albicans (A.Cunn.) Paul G.Wilson subsp. *albicans*

For plants, abbreviation of authors' names follows an internationally agreed standard (Brummitt and Powell 1992), and this publication may be used to set up a checklist, or used for data entry and/or validation checking.

A.Cunn. = Allan Cunningham

L. = Linnaeus

L.f. = Linnaeus filius (son of-)

Sometimes, a space is given between Initial and Surname, others not. It is a matter of preference.

Error checking

Author names as used in Brummit and Powell (1992) could be used to check authors in botanical database. Harvard University also has prepared a downloadable file of botanical authors and made this available on-line². This should prove to be a very valuable file for checking authors' names and dates. Some names databases also include author names (e.g. IPNI 1999, Froese and Bisby 2002). Again Soundex-like techniques as mentioned above could be used to look for similarities between two names. It is the combination of species name and author that is the deciding factor, however, and these are not always easy to check.

If authors are used, then all published names in the database should have an author. In these cases, a Missing Data Values check should be carried out.

Collectors' names

Collectors' names are generally not standardised in collection databases, although standardisation of plant collector's names are being attempted for plant names in the *speciesLink* project in Brazil (Koch, 2003), and at Kew Gardens by Peter Sutton.

Extensive lists of collector's names have been published for some areas, but mainly for botanical collectors (see Steenis -Kruseman 1950, Hepper and Neate 1971, Dorr 1997, Index Herbariorum 1954-1988). There are also a number of on-line resources available:

- Harvard University have recently prepared a downloadable file of botanical collectors and collector teams and made these available on-line.
<http://www.huh.harvard.edu/databases/cms/download.html>
- Index Collectorum – from the University of Göttingen
http://www.sysbot.uni-goettingen.de/index_coll/default.htm
- Directory of Insect Collectors of Southern Africa (Entomological Society of Southern Africa)
<http://www.up.ac.za/academic/entomological-society/collectr/collectr.html>
- Index bio-bibliographicus notorum hominum Nonveilleriana (The Croatian Entomological Society)
<http://www.agr.hr/hed/hrv/bibl/osobe/comentsEN.htm>

There are also a number of hard copy publications, and there are sure to be many more such indexes available in the various disciplines of zoology.

Data entry

It is recommended that names be included in primary species databases in a standard format. The HISPID Standard (Conn 2000) recommends the following:

Primary collector's family name (surname) followed by comma and space (,) then initials (all in uppercase and each separated by fullstops). All initials and first letter of the collector's family name in uppercase. For example, Chambers, P.F.

² <http://www.huh.harvard.edu/databases/cms/download.html>

It is recommended that secondary collectors be placed in a second field. If this is not the case, then it is recommended that they be cited with a comma and space used to separate the multiple collectors. For example:

Tan, F., Jeffreys, R.S.

Where there is a chance of confusion, other given names should be spelt out. For example, to distinguish between Wilson, Paul G. and Wilson, Peter G. (a space is placed after the given name; no punctuation, except as separator between two names, as described above).

Titles should be omitted.

If the family name (surname) consists of a preposition and a substantive, as in many European names (e.g. C.G.G.J. van Steenis), then the preposition is in lower case and the substantive has an initial capital letter. For example:

Steenis, C.G.G.J. van

Other names of similar form include de la Salle, d'Entrecasteaux, van Royen etc. It should be noted, however, that many of these names have been anglicised, particularly in America, such that both parts of the family name are treated as substantive. In such cases, these names can be transferred as follows:

De Nardi, J.C.

The prefixed O', Mac', Mc' and M' (e.g. MacDougal, McKenzie, O'Donnell) should all be treated as part of the substantive and hence transferred as part of the family name. For example:

McKenzie, V.

Hyphenated given names should be transferred as all uppercase, with the first and last initial separated by a hyphen (without spaces), and only the last terminated by a full stop. For example:

Quirico, A-L.

Peng, C-I.

If the collector of the record is unknown, then the term "Anonymous" should be used.

Interpreted information should be enclosed in square brackets, e.g.

Anonymous [? Mueller, F.]

Error checking

As mentioned above, without a standard list of collectors, it is not easy to carry out error checks on collectors' names. This is particularly so in databases that do not follow a standard practice (such as putting surname first as mentioned above). If the database has standardised, then it is quite easy to sort all collector's names in the database and look for slight variations (for example a collector that uses one initial sometimes, and two at others). Extreme care should be taken not to introduce new errors to the database by altering a collector's name without absolute certainty that the change is correct. The initials example, above, is one case where a change could easily introduce new error. Errors that may be correctable are misspellings of surname, for example.

One way to develop a list of collectors is to create a list of unique values from the database in the same way as authority tables are developed for taxon names.

Fields associated with the *Collector-Name* such as *Date-of-Collection* may also be used for error checking. Historians have carried out a considerable amount of work recently on developing itineraries of explorers and collectors, historic scientific expeditions, ship itineraries, etc. Often these are not carried out by scientists, but by historians, and our science can benefit greatly from this work (see resources listed above, along with collections in the libraries (publications, journals, etc.) of many of the world's older museums and herbaria, recent work at the University of California, San Diego, the Scripps Institution of Oceanography, and the National Science Digital Library on capturing and documenting data from cruises as part of the SIO Digital Library Project). Links between those databases and primary species databases can lead to an improvement of both as inconsistencies and errors are detected.

Spatial Data

Spatial location is one of the most crucial aspects in being able to determine the fitness for use of many species-occurrence records. Spatially related biogeographic studies comprise one of the largest uses for these data – studies such as species distributional modelling, biogeographic studies, environmental planning and management, bio-regionalisation studies, reserve selection and conservation planning, and environmental decision support. For a detailed study, see the associated paper on *Uses of Primary Species-Occurrence Data* (Chapman 2005b).

We often think of primary species data as being *point* records of plant or animal occurrences but this is only part of the story. Seldom are collecting locations recorded accurately or precisely enough to be regarded as true points. The accuracy associated with the collection means that the point actually represents an area or a footprint. For example, a location from textual information that says “10 km north of Campinas”, then there is an accuracy associated with the distance of “10 km” (perhaps ± 500 m), an accuracy associated with the direction “north” (i.e. north is somewhere between say NW and NE), and there is an accuracy associated with “Campinas” (is it the city boundary – a polygon – the city centre, etc.). For a more detailed discussion, see Wiczorek 2001a, Wiczorek *et al.*, 2004. In addition, many observational and survey records are recorded from an area (a *polygon*) such as bird observations over a 2 ha area, or within a National Park, or from a regular grid (a *grid*) such as observations from all 10-minute grid squares across Australia (Blakers *et al.* 1984), or from a 10 m X 10 m survey grid, or from along a transect (a *line*) such as a transect survey or records along a road or river (although probably better treated as a polygon derived from buffering the road or river, depending on the scale). See further discussion under *Visualisation of Error* below.

As mentioned previously, a number of programs do exist that can aid in checking and testing for errors in geocodes attached to primary species records. Other tools are available to assist in the original assignment of coordinates to the data from the location information (such as distance and direction from a named location).

The testing of errors in already assigned georeferences involves

- Checking against other information internal to the record itself, for example, State or named district.
- Checking against an external reference using a database – e.g. is the record consistent with the collecting localities of the collector;

- Checking against an external reference using a GIS, including “point-in-polygon” tests – that the record falls on land rather than at sea, for example;
- Checking for outliers in geographic space; or
- Checking for outliers in environmental space.

Data Entry and Georeferencing

As stressed throughout these documents, error prevention is preferable to error detection, and the georeferencing or geocoding of records is one of the greatest sources of error in the databasing of species-occurrence data. Many new tools are now being developed to assist with the process of adding coordinates (especially latitude and longitude) to primary species data. This is not an easy process, however, especially as much of the legacy data (early collections in museums and herbaria collected over the past 300 or 400 years) carry little geographic information other than a general description of the location where they were collected (Chapman and Milne 1998). These collections were often made before modern settlements were built and named, and before roads were built. Many were collected from horseback or by boat, days from the last settlement and reference points were often difficult to determine. Many of the reference points no longer occur on modern maps and, in many cases where they do occur, they are ambiguous. Where geocodes are given, they are often not very accurate (Chapman 1999) and have generally been added at a later date (*retrospective georeferencing* – Blum 2001) by those other than the collector (Chapman 1992).

Definitions:

Before proceeding, there are a number of terms whose use in this document need definition. Some of these terms are used differently elsewhere, and different disciplines use different terms to define similar processes.

Geocode: As used in this paper, a geocode is the code (usually an x, y coordinate) that records the spatial positioning of a record according to a standard reference system. Here it is used to record a position on the surface of the earth. For species-occurrence data, the geocode is given according to one of several standard geographic reference systems with Universal Transverse Mercator, and latitude and longitude being two of the more common, and may be recorded in one of a number of ways (meters; decimal degrees; degrees, minutes, seconds; degrees, decimal minutes, etc.). Definitions of the term geocode are broad and wide-ranging. In many GIS applications it refers to an address and Zip Code, in marketing terms it refers to a demographic characterisation of a neighbourhood, and in some cases (Clarke 2002) it refers only to the location in “computer readable form”. Also sometimes called a *georeference* or *coordinate*.

Georeferencing: In this paper georeferencing is used to describe the process of assigning geographic coordinates to a record that links it to a geographic location on earth. Also sometimes called *geocoding*.

Database design

The design of databases for primary species-occurrence data should ensure that there are fields to properly cater for information that is often wrongly placed in the locality field – data such as habitat and habit information and geographic notes. An example of a distribution field with mixed information (from Fishbase³ for *Perca fluviatilis*) is:

³ <http://www.fishbase.org/>

“Throughout Europe and Siberia to Kolyma River, but not in Spain, Italy or Greece; widely introduced. Several countries report adverse ecological impact after introduction”.

Such mixed fields are very difficult to treat in an automated way using parsing algorithms and are not consistent with the philosophy and design of relational databases where the information can be stored in Memo fields.

There are several additional fields that can be added to a species-occurrence database to assist in data cleaning and that can lead to an improvement in documenting data quality. Such fields include:

- ***Spatial accuracy*** – a field that records (preferable in meters, but sometimes in coded form) the accuracy with which a record’s location has been determined.
- ***Named Place, Distance and Direction*** – some databases include “Nearest Named Place”, “Distance” and “Direction” in separate fields as well as a plain text locality field. The inclusion of such fields can aid in geocode determination as well as in error checking.
- ***Geocode method*** – a field (or fields) that records how the geocode was determined – may include (Chapman 2005a)
 - use of differential GPS;
 - handheld GPS corrupted by Selective Availability (i.e. a recording prior to 1 May 2000);
 - A map reference at 1:100 000 and obtained by triangulation using readily identifiable features;
 - A map reference using dead reckoning;
 - A map reference obtained remotely (eg. in a helicopter);
 - Obtained automatically using geo-referencing software using point-radius method;
 - Obtained from database using previously georeferenced locality.
- ***Geocode type*** – records the type of locality description that was used to determine the geocode.

In a paper on the point-radius method of georeferencing locality descriptions, Wieczorek and others (2004) provide a table of nine types of locality descriptions found in natural history collections. The first three of these they recommend should not be georeferenced, but an annotation be given as to why it was not georeferenced. Some databases use a centroid with a huge accuracy figure (e.g. 100,000,000 meters). This has the drawback of users extracting the data only using the geocode and not the associated accuracy field and ending up with what looks like a point without its associated huge radius. The Wieczorek method overcomes this drawback by not providing such a misleading geocode. The nine categories listed by Wieczorek *et al.* (2004) are (with modified examples):

1. ***Dubious*** (e.g. ‘Isla Boca Brava?’)
2. ***Cannot be located*** (e.g. ‘Mexico’, ‘locality not recorded’)
3. ***Demonstrably inaccurate*** (e.g. contains contradictory statements)
4. ***Coordinates*** (e.g. with latitude or longitude, UTM coordinates)
5. ***Named place*** (e.g. ‘Alice Springs’)
6. ***Offset*** (e.g. ‘5 km outside Calgary’)

7. **Offset along a path** (e.g. ‘24 km N of Toowoomba along Darling Downs Hwy’)
8. **Offset in orthogonal directions** (e.g. ‘6 km N and 4 km W of Welna’)
9. **Offset at a heading** (e.g. 50 km NE of ‘Mombasa’)

Each of these would require a different method of calculation of the accuracy as discussed in the paper (Wieczorek *et al.* 2004)

Georeferencing Guidelines

Two excellent guidelines have been developed to assist data managers with georeferencing. The Georeferencing Guidelines developed by John Wieczorek at the Museum of Vertebrate Zoology in Berkeley (Wieczorek 2001) and the MaPSTeDI (Mountains and Plains Spatio-Temporal Database Informatics Initiative) guidelines (University of Colorado 2003) are two of the most comprehensive. I understand that there are also guidelines developed by Conabio in Mexico (CONABIO 2005), which are being translated into English, and thus will soon be available in both Spanish and English.

Edit controls

Edit controls involve business rules that determine the permitted values for a particular field. One of the most frequent errors in spatial databases is the accidental omission of the ‘–’ (minus) sign in records from the southern or eastern hemispheres. If the database is a database of all southern hemisphere records (a database of Australian records for example), then it should be automatic that all records are given a “negative” latitude. Databases of mixed records are, of course, more difficult to deal with, but the country and state fields could be used to check against the latitude or longitude.

Not all databases are set up correctly initially, and this can allow errors that should never occur. For example, latitudes greater than 90° or less than –90° and longitudes greater than 180° or less than –180°. If these are permitted by the database, then the database needs to be modified, otherwise checks need to be run on a regular basis to ensure that errors like these do not occur and are corrected.

Using existing databased records to determine geocodes

Information already included in the database can be used to assign georeferences to new records being added. A simple report procedure can be incorporated that allows for a search to ascertain if a specimen from the same locality has already been databased and assigned a geocode.

For example, you are about to database a collection that has the location information “10 km NW of Campinas” but no georeferencing information. You can search the database for “Campinas” and look through the collections already databased to see if a geocode has already been assigned to another collection from “10 km NW of Campinas”. This process can be made a lot simpler if the database structure includes fields for “Nearest Named Place”, “Distance” and “Direction” or similar, in addition to the traditional free text locality description.

This methodology has the drawback that if the first geocode had been assigned with an error, then that error will be perpetuated throughout the database. It does, however, allow for a global correction if such an error is found in any one of the collections so databased. If such a method is used to determine the geocode it should be so documented in the *Geocode method* field (see above).

With linked databases, such as the Australian Virtual Herbarium (CHAH 2002), *speciesLink* (CRIA 2002), or the GBIF Portal (GBIF 2004), on-line procedures could be set up to allow for a collaborative geocoding history to be developed and used in a similar way. Such collaboration may be carried out through the use of Web Services (Beaman *et al.* 2004, Hobern and Saarenmaa 2005). Of course, one drawback of this is that there is a certain amount of loss of control within your database, and an error in another database can be inadvertently copied through to your own database. Where this is done then the source-id should be attached to the record so that later updates and corrections can be incorporated. Good feed back mechanisms would need to be developed between institutions to ensure that, firstly errors were not perpetuated inadvertently, and secondly that information on errors that are detected are fed back to the originating database as well as other dependent databases.

Many plant collections are distributed as ‘duplicates’ to other collection institutions. Traditionally this has been done prior to georeferencing, and one can often find exactly the same collection in a number of different institutions, all with different georeferencing information. To circumvent these discrepancies, geocodes need to be added before distribution, or a collaborative arrangement entered into between institutions. As explained earlier, it costs a lot in both time and money to add geocodes, it is an extremely wasteful exercise if several institutions individually spend time and resources georeferencing the same collections. The waste is further compounded if different geocodes are given to the same collection in those separate institutions.

Automated geocode assignment

Automated georeferencing tools are based on determining a latitude and longitude from the textual locality information using a distance and direction from a known location. Ideally, databases include at least a “Nearest Named Place”, “Distance” and “Direction”, or better still, “Named Place 1”, “Dist 1”, “Dir. 1”, “Named Place 2”, “Dist 2”, “Dir 2”. Thus “5 km E of Smithtown, 20 km NNW of Jonestown” would be appropriately passed into the six fields cited above.

As most databases are not so structured, attempts are being made to develop automated parsing software to parse free-text locality descriptions into basic “Nearest Named Place”, “Distance” and “Direction” fields, and then using these fields, in association with appropriate Gazetteers to determine the georeference (see *BioGeomancer* below). At the same time as the geocode is determined in this way, the Geocode Accuracy should be recorded in an extra field and where possible, the results checked by experts against the original to avoid unanticipated errors. In any case, such parsing should not in any way tamper with the original “Locality” data (field), but be additional information added. It can thus always be used to check the accuracy of the parsing exercise.

Drawbacks of this methodology include possible errors in the Gazetteers (most publicly available gazetteers have a considerable number of errors (see for example, figure 15), Nearest Named Place locations may refer to quite a large area (see comments below on assigning accuracy), many location fields are not as straight forward as those cited above, often historic place names are used, and many distances on collection labels are “by road” distances rather than direct, although this is seldom stated on the label itself. Accuracy fields need to take into consideration these issues as well as the error inherent in vector distances – does “South West” mean between “South” and “West” or between SSW and WSW. As this distance from the source increases, the inherent error in these will also rapidly increase (see discussion in Wieczorek *et al.* 2004). The use of this method in conjunction with a simple

GIS would provide the opportunity for the operator to see the record on a map and to then “grab and drag” the point to a more appropriate place – for example to the nearest road.

Geocoding software

A number of on-line and stand-alone tools have been developed to assist users with georeferencing their collections. Three are mentioned here – two ‘on-line’ and two ‘stand-alone’.

BioGeoMancer

BioGeoMancer is an automated georeferencing system for natural history collections (Wieczorek and Beaman 2002). In its present state, BioGeoMancer is a prototype system, and the comments below do not consider planned enhancements that are sure to improve its useability. BioGeoMancer can parse English language place name descriptions and provide a set of latitude and longitude coordinates associated with that description. The parsing of free-text, English language locality data provides an output of nearest named place, distance and direction, in the format (Wieczorek 2001b):

- 2.4 km WNW of Pandemonium
- Springfield, 22 miles E
- Springfield, 0.5 mi. E of Pandemonium

Like a number of other programs (e.g. Diva-GIS, eGaz) it takes the parsed information and in conjunction with an appropriate gazetteer, calculates a set of latitude and longitude coordinates. BioGeoMancer has the advantage over other geocoding programs in that it provides the parsing of the text. It is the first such geo-parsing program available to the public and researchers over the internet.

The image shows a screenshot of the BioGeoMancer web interface. At the top left is the Peabody Museum logo with a bird illustration and the text 'Peabody Museum NATURAL HISTORY'. To the right is the 'BioGeoMancer' title in large blue letters, with navigation links for 'Home', 'Documentation', 'Batch forms', and 'Partners'. The main heading is 'Georeference a single locality'. Below this are four input fields: 'Country', 'State or province', 'Admin Level 2', and 'Locality'. To the right of the 'Admin Level 2' field is a note: 'e.g., county, shire, municipio. Leave blank if not known.' Below the input fields is a 'Format results as:' section with radio buttons for 'html' (selected), 'map', and 'xml'. At the bottom center is a 'Submit Query' button.


Fig. 2. Single locality BioGeoMancer query form <http://biogeomancer.org/> (Peabody Museum n.dat.).

The BioGeoMancer program exists in two forms. The first is a single specimen Web query form (figure 2) that allows the user to type in a locality and get a georeference returned. The second form, a batch process, accepts data through either an HTTP/CGI interface in a comma-delimited version (figure 3) or in a SOAP/XML version and provides a return file with georeferenced records either in delimited, html, table (figure 4), or xml format. This

project has recently received a considerable boost in funding and expanded to become a worldwide collaboration attempting to develop new and improved georeferencing tools.

Batch-mode automated georeferencing for natural history collections

HTML Prototype for worldwide locations



```

"12931","Mexico","Veracruz","","12 km NW of Catemaco"
"12932","Mexico","Veracruz","","6 km SW of San Andres Tuxtla"
"13158","USA","Florida","","Sound off Captiva Pass"
14061 USA FL Clearwater Bay"
"15938","USA","FL","","0.24 mi. N of Micanopy; 10 mi S of Gainesville"
"56508","Australia","","","2 miles W of Leura"
"60368","Australia","","","12 km N of Lake Cargelligo"
"105653","Mexico","Oaxaca","","Monte Alban"
"136079","USA","SC","","8 mi NE of Charleston"
"136319","Malaysia","","","Kinabalu South; 7.5 km NNE of Tenompok"
"136341","USA","TX","","Redfish Point; Copano Bay"
"136364","USA","TX","","0.25 mi N of Lap Reef Pass: Copano Bay"
"136491","USA","FL","","Clearwater Beach"
"211939","USA","NC","","16 mi NW of Marion"
"48656","USA","FL","","Fernandina Beach"
"48657","USA","FL","Levy","Hog Island"

```

Fig. 3. Input format for the BioGeoMancer web-based Batch-mode automated georeferencing tool for natural history collections <http://georef.peabody.yale.edu/yu/bgm-forms/batch-int02.html> (Peabody Museum n.dat.).

Biogeomancer Results										
Summary										
Query Id	Query Country	Query Adm1	Query Adm2	Query Locality	Number of records matched	Centroid Latitude	Centroid Longitude	Error radius (km)	Multipoint match	Bounding Box
12931	Mexico	Veracruz		12 km NW of Catemaco	1	18.49331	-95.19701	0.0	MULTIPOINT(-95.19701 18.49331)	BOX(-95.19701 18.49331, -95.19701 18.49331)
12932	Mexico	Veracruz		6 km SW of San Andres Tuxtla	1	18.41167	-95.25682	0.0	MULTIPOINT(-95.25682 18.41167)	BOX(-95.25682 18.41167, -95.25682 18.41167)
13158	USA	Florida		Sound off Captiva Pass	1	26.60917	-82.22222	0.0	MULTIPOINT(-82.22222 26.60917)	BOX(-82.22222 26.60917, -82.22222 26.60917)
14061	USA	FL		Clearwater Bay	1	27.97222	-82.82083	0.0	MULTIPOINT(-82.82083 27.97222)	BOX(-82.82083 27.97222, -82.82083 27.97222)
15938	USA	FL		0.24 mi. N of Micanopy, 10 mi S of Gainesville	1	29.50614	-82.325	0.0	MULTIPOINT(-82.32500 29.50614)	BOX(-82.32500 29.50614, -82.32500 29.50614)
56508	Australia			2 miles W of Leura	2	-28.449995	149.925235	587.4	MULTIPOINT(149.55188 -23.18333, 150.29859 -33.71666)	BOX(149.55188 -33.71666, 150.29859 -23.18333)

Fig. 4. Sample partial output in tabular form from the BioGeoMancer web-based Batch-mode automated georeferencing tool for natural history collections (Peabody Museum n.dat.).

Where more than one option is possible, then all are reported under that ID. Where no options are obvious, then the record is not returned. The *Bounding Box* column provides the calculated accuracy. The system works well for a lot of data, but does have difficulty with text that is not easily parsed into the above named place, distance and direction. Other noted

issues in the current version include (future enhancements are planned that will reduce these):

- It is restricted to English-language descriptions.
- Accuracy is reported only as a bounding box in the present version, and this could be improved. Already, a related program developed by John Wieczorek (2001b) – the Georeferencing Calculator - can supply this information <http://manisnet.org/manis/gc.html> and this is likely to be linked to BioGeoMancer at a later date. Already work has begun on a method of assigning accuracy automatically through what has been termed the “point-radius method” for georeferencing and calculating associated uncertainty (Wieczorek *et al.* 2004)
- Two named localities (e.g. “10 km W of Toowoomba toward Dalby”) produces a null result.

Another parsing program, RapidMap Geocoder (NMNH 1993) was developed in 1993 by the US National Museum of Natural History and the Bernice P. Bishop Museum in Hawaii, for use only with Hawaiian localities. However it was not considered successful and was discontinued. Some useful information on the parsing methodologies used, however, is available on the internet at: http://users.ca.astound.net/specht/rm/tr_place.htm.

GeoLoc-CRIA

GeoLoc is a simple web-based program for finding localities in Brazil, a known distance and direction from a gazetted locality. It has been developed at CRIA (Marino *et al.* in prep.). GeoLoc works in a similar way to the eGaz program (see below) and can be found at <http://splink.cria.org.br/tools/> (CRIA 2004a). The prototype includes a number of gazetteers and provides the user with the potential to select which gazetteer if more than one is available for an area, and also provides a calculated error value.

An example can be seen in figure 5, where the latitude and longitude of a locality 25 km NE of Campinas in São Paulo, Brazil is sought. Firstly one finds the locality for Campinas using one of a number of Gazetteers, or the *speciesLink* records (records obtained through distributed searching of a range of databases mainly in the State of São Paulo). Then by adding “25 km” and “NE” (circled) and clicking on the relevant ‘Campinas’ (arrow), the results will appear on an associated map (figure 6). The geocode is given as -46.9244, -22.7455 with an error of 9.754 km (circled). This information (latitude, longitude and error) are already stored in the Microsoft paste buffer and can be pasted into any Microsoft Windows compatible file such as Word, Excel, and Access. The map also shows the location of “Campinas” from the three sources – the one in red being the one chosen, along with the point “25 km NE of Campinas”. The map can be zoomed and panned, and various environmental layers turned on or off.

The program can also link to an EXCEL spreadsheet of localities and produce an html table of results for further searching, or an EXCEL spreadsheet. The main drawback of the program is that it is only available for use with Brazilian locations. The algorithms are currently being incorporated into the wider Biogeomancer project.

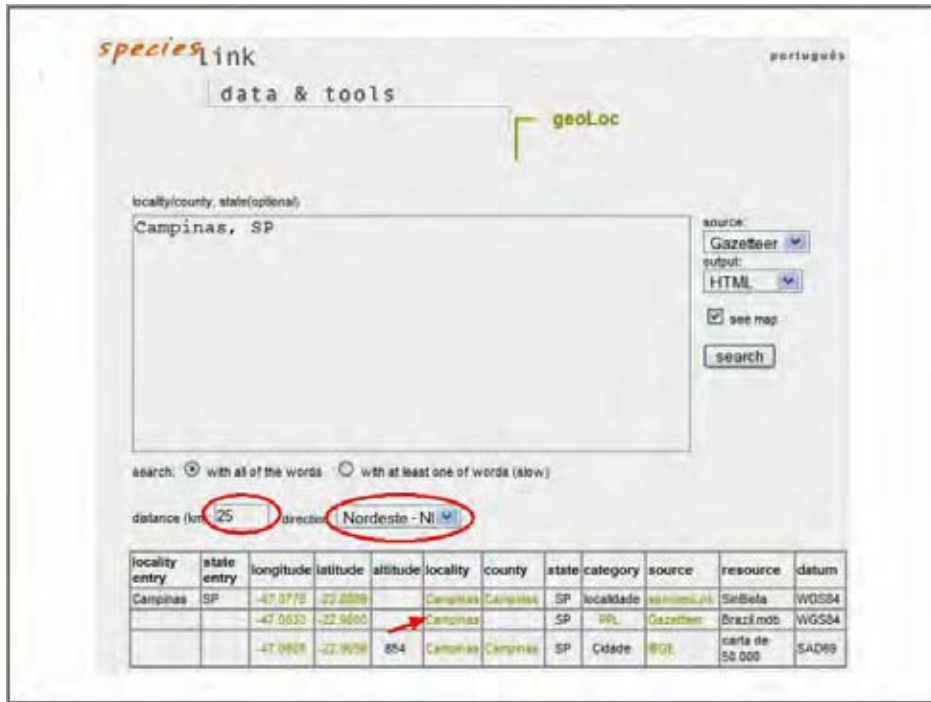


Fig. 5. Using CRIA's 'geoLoc' program to find the geocode for a locality 25 km NE of Campinas, SP.



Fig. 6. Results of the above selection showing the location of "Campinas" (from the various sources) and the point 25 km NE of Campinas, with associated geocode information and error (circled)

GEOLocate

GEOLocate (Rios and Bart *n.dat.*) is a georeferencing program developed by Tulane University's Museum of Natural History and is designed to facilitate the task of assigning geographic coordinates to the locality data associated with natural history collections. The primary goals of GEOLocate are to:

- develop an algorithm to convert textual natural history data into latitude and longitude for North America;
- provide an interface for visualisation and further adjustment of generated coordinates;
- provide a simple solution for users to import and georeference their data;
- provide an auto-updating feature.

The algorithm first standardises the locality string into common terms and parses out distances, direction and key geographic identifiers such as the named place. This information is then used in conjunction with gazetteers (including placenames, river miles, landuse and road/river crossing data) to determine the geographic coordinates. The program also allows the user to “snap” localities to the nearest water-body.

The program is available from the University of Tulane, and an on-line demonstration is available at: <http://www.museum.tulane.edu/geolocate/demo.aspx> (figure 7).

GEOLocate
Georeferencing Software for Natural History Collections

Online Demo

Locality String: Cambridge

Country: Canada (Required)
State: ON (Required for U.S.)
County: (Required for U.S.)

Options (require county data, US only):
 Enable Water Body Matching
 Enable Hwy/River Crossing Detection

[Instructions](#)

2 possible location(s) found

Latitude	Longitude
43.383333	-80.316667
45.2	-64.116667

Click map to:
 Zoom In
 Zoom Out
 Recenter
 Adjust Marker

Marker Coordinates:
Latitude: 43.383333
Longitude: -80.316667

Fig. 7. An example of the GEOLocate interface using the on-line demo version to identify the geographic coordinates for Cambridge, Ontario.

The program only works for North America (Mexico, USA and Canada), but the developers are currently working on extending it to include the entire world. Other developments will

include DiGIR compatibility, multi-lingual support, and advanced validation techniques (N.Rios pers. com. 2004).

eGaz

eGaz (Shattuck 1997) is a program developed at the CSIRO's Australian National Insect Collection to assist museums and herbaria to identify and add geocodes to their specimen records. With the development of the data entry and specimen management software, BioLink (Shattuck and Fitzsimmons 2000), it was incorporated into that software package. eGaz is available as part of the Biolink package (see below).

eGaz eliminates the need for paper based maps and rulers to determine the latitude and longitude for cities, towns, mountains, lakes and other named places. eGaz can also calculate latitude and longitude for sites a known distance and direction from a named place. The program allows for the easy inclusion of Gazetteers from any region, and Gazetteers for much of the world are available for download from the CSIRO site (<http://www.biolink.csiro.au/gazfiles.html>).

eGaz is a Microsoft Windows based product that provides two windows, a Gazetteer window and a Map window (figure 8). It allows the user with a location in the form of a "Named Place", "Distance" and "Direction" to obtain a geocode and transfer that to a file.

The example shown in figure 8 is of obtaining the latitude and longitude of a position "80 km SW of Toowoomba", Queensland, Australia. The first step is to load the appropriate Gazetteer and select "Toowoomba" from it (**A**). There are a number of options, but I have selected the Toowoomba City (labelled POPL for Populated Place). The location of Toowoomba appears on the map in red (**B**). The distance "80" is typed into the Distance field and the pull down menus used to select "km" and "SSW" (**C**). The selected location appears on the map as a blue dot (**D**). The location, along with the latitude and longitude also appears on the bottom of the Gazetteer window (**E**). By right clicking on this area and selecting "Copy" that information can be copied and pasted into any Microsoft Windows compatible file (Word, Excel, Access). The Latitude and Longitude (to 1 arc-minute resolution) also appears (**F**), and this can similarly be copied to a file. Alternatively, by going to the Edit menu and select "Copy Lat/Long" the geocode can be copied to an accuracy of one arc-second.

One can also go to the map itself and zoom in to the point. Other layers such as a road network (in ESRI Shape file format) can be loaded to allow more accurate positioning of the point – i.e. perhaps move it to the nearest road if collecting was done from a vehicle. The selection tool can then be used to click on the point to obtain the geocode to one arc-second resolution. Again right clicking with the mouse, or using Edit/Copy Lat/Long, allows the information to be copied to an appropriate file.

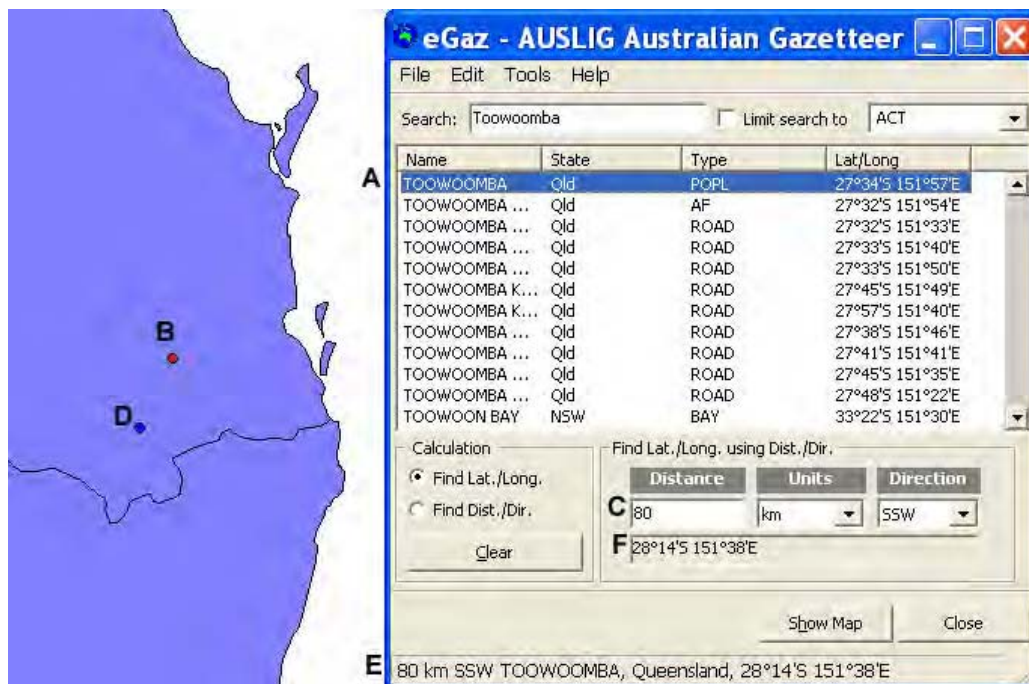


Fig. 8. Sample output from eGaz, showing the determination of latitude and longitude for a position 80 km SSW of Toowoomba, Queensland, Australia. **A.** Information on Toowoomba from Gazetteer. **B.** Mapped location of Toowoomba. **C.** Input showing 80 km SSW of highlighted location. **D.** Mapped location 80 km SSW of Toowoomba. **E.** Details on location. **F.** Latitude and Longitude of new location.

Diva-GIS

Diva-GIS is a free GIS program developed for use in museums and herbaria. It includes an algorithm that assists in assigning coordinates to specimen data where this is lacking. Some pre-processing is necessary to organise the data into a format acceptable to the program, but a number of databases are already beginning to structure their data in this way. The input file requires the textual location data to be parsed into a number of specialised fields. These are “Named Place1”, “Distance 1”, “Direction 1” and “Named Place2”, “Distance 2”, “Direction 2”. For example the locality record:

“growing at a local place called Ulta, 25.2 km E of Chilla”

would be parsed to:

Named place 1:	Ulta
Distance 1:	
Direction 1:	
Named Place 2:	Chilla
Distance 2:	25.2km
Direction 2:	E

and

“14 km ESE of Sucre on road to Zudanez” would parse to:

Named place 1:	Sucre
Distance 1:	14 km
Direction 1:	ESE
Named Place 2:	Zudanez
Distance 2:	
Direction 2:	

Just one set of “Named Place”, “Distance” and “Direction”, however, will be able to provide the geocoding for many records, and this is all the information most institutions will have. The authors of the Diva-GIS (Hijmans *et al.* 2005) recommend rounding the distance down to whole numbers to account for inaccuracies in the data, and to cater for cases where 25 km North of a place, really means 25 km North by road and not in a direct line. I would recommend to the contrary, and would record the most accurate figure given, and place an accuracy figure in an “Accuracy” field in meters.

Once an input file has been selected, an output file named, and the appropriate field names selected from a pull-down list, the algorithm is run and produces an output file (figure 9). The algorithm uses an appropriate Gazetteer to assign coordinates.

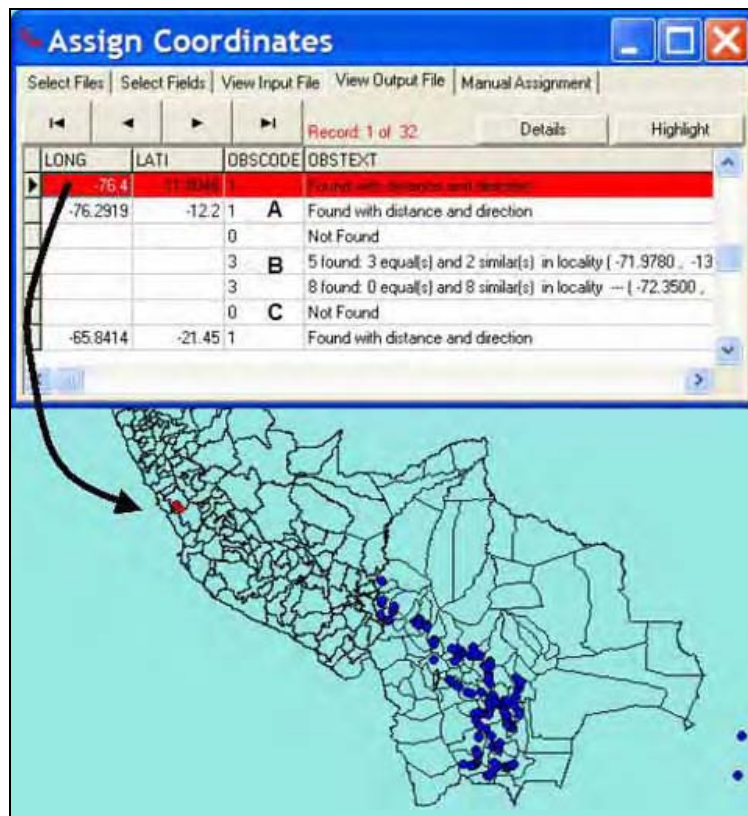


Fig. 9. Results from Diva-GIS showing point records with geocodes automatically assigned. **A.** Unambiguous geocodes found by the program and assigned. **B.** Ambiguous geocodes identified. **C.** Appropriate geocodes not found.

As shown in the example (figure 9), the program has found unambiguous matches in the Gazetteer(s) for a number or records using the “Named Place” field in the input file and assigned those records an appropriately calculated geocode (A). Once the output file has been loaded and a shape file created, each of these records can be highlighted to produce a flashing

point on the map. In a number of other cases, the program has found several possible matches in the Gazetteer(s) for the “Named Place” and reported on that appropriately (**B**). In yet other cases (**C**) the program has been unable to find a match in the Gazetteer.

In the case of records where a number of possible matches were found, one can go to the next stage by double clicking on one of the (**B**) records and producing another output file (figure 10).

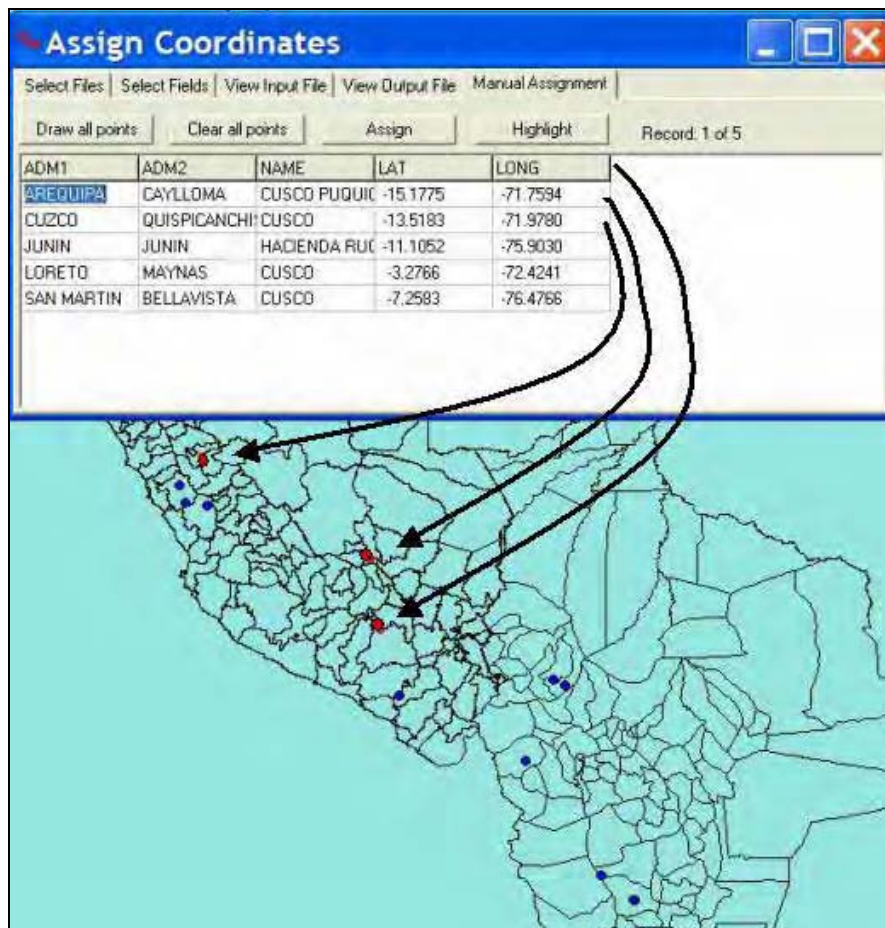


Fig. 10. Results from Diva-GIS showing alternate geocodes for a record where use of the Gazetteer has produced a number of credible alternatives.

In the case of the record shown in figure 11, the program has identified five possible alternative locations from the Gazetteer(s) and presented these alternatives on the GIS for the user to choose. When one is chosen, it is just a matter of clicking on the “Assign” button for that to be assigned to the output file. Alternatively, one can decide on another location altogether and use the “Manual Assignment” to add a geocode or modify one of the assigned ones.

Geocode checking and validation

There are four main methods that can be used for checking and validating geocodes on specimen records once databased. These are the use of databases for checking internal inconsistencies, the use of geographic information systems, the use of environmental space to check for outliers and the use of statistics to check for outliers in geographic or environmental space.

Using Databases

Internal checks

Most species and species-related databases include a certain amount of redundant information. For example, the State in which the collection was made as well as a field for textual location information. Some databases also include a “nearest named place” and this may also duplicate information within the locality field. Checks can then be made to check that the cited town or nearest named place in one field, is located within the correct State or district, or even country as cited in another field.

Checking information in a database between similar records is also possible, for example, checking all localities against the supplied latitude and longitude. One may have a database with 5 collections from one location – “10 km N of Campinas, SP” for example. Do they all have the same latitude and longitude or are one or more significantly different to the others? See also discussion on *Ordinal Association Rules* below.

Data Cleaning (speciesLink)

CRIA’s Data Cleaning module of the speciesLink Distributed Information System (CRIA 2002) includes a number of routines for identifying possible errors to help collection managers in processing their data. At the moment this is only in Portuguese, but an English version is proposed. One portion of this tool is to identify errors in names. Routines include:

- Listing of all names (family, genus, species, subspecies) along with the number of occurrences in the databases accessed. A brief look at one example (figure 11) shows a number of obvious problems. The first line shows that there are 101 occurrences in the database with records not identified at any level from family below. The second line shows one occurrence with a family name “4606euphorbiaceae”, and line 3 shows 5 records in Acanthaceae identified to family only.

family	genus	species	subspecies	total
[]	[]	[]	[]	101
[4606euphorbiaceae]	<i>sp</i> [Julocroton]	[humilis]	[var. subpannosus]	1
[Acanthaceae]	[]	[]	[]	5

Fig. 11. Extract from CRIA Data Cleaning module showing some possible errors.

- Examining possible errors in generic names. This is where the family name is the same, the species name is the same, the generic names are similar (identified using soundex-like algorithms) but the spelling of the generic name is different. This output also shows the number of occurrences of each in the database being studied, and the total number of occurrences in all databases accessed through speciesLink. The example (figure 12) shows two different spellings of the genus *Hieronyma* (along with two spellings of *alchornioides*, but those are identified under a different routine) along with the number of occurrences of each. One can click on the “*sp*” and it takes you to a search of a range of databases both internal to the organisation as well and external, and includes such resources as the International Plant Name Index (IPNI), species 2000, etc. which can all help the user determine which may be the correct spelling.

family	genus	species	subspecies	ocor_col	ocor_total
[Euphorbiaceae]	SP [Hyeronima]	[alchorrneoides]	[]	3	3
[Euphorbiaceae]	SP [Hieronyma]	[alchorrneoides]	[]	1	1
[Euphorbiaceae]	SP [Hyeronima]	[alchornioides]	[]	9	9
[Euphorbiaceae]	SP [Hieronyma]	[alchornioides]	[]	17	17

Fig. 12. Extract from CRIA Data Cleaning module showing some possible errors.

- Examining possible errors in species names or epithets. Like the generic names, this looks for names where the genus is the same, the soundex for the species epithet is the same, but there is a difference in the spelling of the species epithet. Again the output shows the total number of occurrences in the database being studied, the total occurrences in all databases accessed and the status of the name in species2000⁴. The example (figure 13) shows a number of species names with alternatives. The number of occurrences of each name along with the status in species2000 if available can give an indication of which of the spellings may be an error.

genus	species	subspecies	ocor_col	ocor_total	status_sp2000
SP [Acacia]	[polyphylla]	[]	83	217	accepted name
SP [Acacia]	[polyphyllla]	[]	1	1	
SP [Banisteriopsis]	[argyrophylla]	[]	25	164	
SP [Banisteriopsis]	[argirophylla]	[]	2	2	
SP [Bauhinia]	[cuyabensis]	[]	19	19	provisionally accepted name
SP [Bauhinia]	[cuiabensis]	[]	1	2	
SP [Bignonia]	[unquis-cati]	[]	1	5	unambiguous synonym
SP [Bignonia]	[unquiscati]	[]	2	2	

Fig. 13. Extract from CRIA Data Cleaning module showing some possible errors.

- Examining differences and possible errors in author names. Figure 14 shows the number of possibilities for just one species name. Again clicking on the “**SP**” a search of other databases can be carried out to help determine which may be the best alternative to use.

genus	species	subspecies	author	ocor_col	ocor_total
SP [Actinostemon]	[concolor]	[]	[Mull.Arg.]	0	1
SP [Actinostemon]	[concolor]	[]	[(Spreng.) N.Arg.]	0	15
SP [Actinostemon]	[concolor]	[]	[(Spreng.) Huell.Arg.]	0	17
SP [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	1	2
SP [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	0	2
SP [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	0	52
SP [Actinostemon]	[concolor]	[]	[(Spreng.) Mull.Arg.]	85	139
SP [Actinostemon]	[concolor]	[]	[(Spreng) Mull. Arg.]	0	6
SP [Actinostemon]	[concolor]	[]	[(Spr.) Huell. Arg.]	0	1
SP [Actinostemon]	[concolor]	[]	[(Spr.) Huell.Arg.]	0	2

Fig. 14. Extract from CRIA Data Cleaning module showing some possible errors.

- Examining differences in family names and in subspecies names works in a similar manner.

Other routines are used to identify possible geographic errors in the datasets, and these are treated under *Spatial Data* below. CRIA is not a custodian of the data, and makes no changes to the data, but provides a service to data custodians, to help them identify possible errors in their databases. It is then up to the custodians to decide which are the correct formats and which records should be corrected and which not.

⁴ <http://www.species2000.org>

External databases

By linking to external databases, errors in various aspects of the species-occurrence data can be identified. Such databases can include Digital Elevation Models, spatial topographic databases, gazetteers and collector's itineraries.

More sophisticated databases can be used to check the accuracy of the altitude fields by comparing the altitude cited with that of a databased Digital Elevation Model (DEM). It is important that the DEM used be at an appropriate scale, and due to the varying accuracy of most specimen data, can lead to false or misleading errors if not used critically. Such a technique has been used successfully in ERIN (Environmental Resources Information Network) in Australia for over 10 years (Chapman unpublished). The process uses batch processing using an ORACLE[®] database and can check (or assign) altitude records to over 3000 records a minute.

More recently, sophisticated spatial databases have been developed such as ESRI's Spatial Database Engine (ArcSDE[®]) (ESRI 2003) and PostGIS that allow for more complicated database searching using the geocodes themselves. This type of software, however, is very expensive, and very few museums or herbaria are likely to afford them or have the need for them and for that reason, these methods are not further outlined further in this paper.

Gazetteers exist for most of the world in one form or another, and frequently these are available as a downloadable database. They can be used to check appropriate fields within the specimen database for accuracy. Care needs to be exercised with the use of many of these databases as often they, themselves, contain errors (see for example figure 15), and it is important that the right gazetteer for the area, at an appropriate scale is used. Also, many named place names may be ambiguous (e.g. there are hundreds of "Sandy Creek"s in Australia) (Chapman and Busby 1994), or involve historic place names that do not occur in the modern gazetteer. There is also the issue of what a place name may actually mean (Wieczorek 2001a). One of the aspects of the new BioGeomancer project (see comments elsewhere) is the integration of Gazetteers with biological databases using Web Service technology. It is also hoped to improve gazetteers through public participation, and to especially begin including historic collection locations.

One method that is seldom used, but that has great potential, is cross checking against databases of collectors' localities. To date, very few such databases exist (but the Harvard database referenced above would be a good general starting point for botany⁵), and others are gradually being developed. Peterson *et al.* (2003) recently suggested a novel statistical method using the birds of Mexico as an example. They ordered the collections of a particular collector in temporal order and for each day (or group of days) impose a maximum radius of likely movement. Using a formula-based approach in EXCEL, they identified possible errors in specimens that fall outside the calculated range. Similar methods to this could be carried out in the database itself – see discussion under *Ordinal Association Rules*, below. Such a method will only work, however, if the databased collections from the collector are large enough to create such an itinerary.

⁵ <http://www.huh.harvard.edu/databases/cms/download.html>

ii. GIS Checks

Geographic Information Systems (GIS) are very powerful tools that have become much more user friendly in recent times. GISs range from expensive, high functionality systems to free, off-the-shelf products with more limited functionality. Many of the free GISs are powerful enough, however, to provide much of the functionality required by a herbarium or museum, and can be easily adapted to provide a range of data checking and data cleaning routines.

	Points	Lines	Polygons
Points	<ul style="list-style-type: none"> ▪ is a neighbour of ▪ is allocated to 	<ul style="list-style-type: none"> ▪ is near to ▪ lies on 	<ul style="list-style-type: none"> ▪ is a centroid of ▪ is within
Lines		<ul style="list-style-type: none"> ▪ crosses ▪ joins 	<ul style="list-style-type: none"> ▪ intersects ▪ is a boundary of
Polygons			<ul style="list-style-type: none"> ▪ is overlain by ▪ is adjacent to

Table 3: *Relationships between classes of objects (from Gatrell 1991)*

The GIS can also be used to check for logical consistency within the database. Redundancy in topological encoding can be used to detect flaws in data structure such as missing data and unlabelled polygons (Chrisman 1991). GIS allows the interrelation of spatial layers to detect errors and that, along with visualisation, is its major strength.

The use of a simple GIS to plot points (specimen records) against polygons (regions, States, Countries, soils) can aid in detecting mismatches in the data (either geographic or altitudinal). This is a common test used in GIS systems and is known as the “point-in-polygon” method – it is used in GIS to make sure marine buoys don’t occur on land, that rivers don’t occur outside their flood plains, etc. One of the most important tests a GIS can perform on primary species data is to check that records that are supposed to be on the land actually are on land, and those that are supposed to be in the ocean, are. It is obvious when one first loads a large data set into a GIS that many records are obviously in the wrong place just from this simple check. Checks for misplaced records using a GIS can range from simple visual inspection to more automated checking. Visual inspection can also be valuable in determining if records fall in the correct country, for example. If you have a database of records from Brazil, by using a GIS you can quickly identify records that are misplaced in such a way that they are outside of Brazil. For example, in figure 15, records from a publicly available Gazetteer of Brazilian place names have some obvious errors. Errors in specimen records can similarly be identified using this methodology.

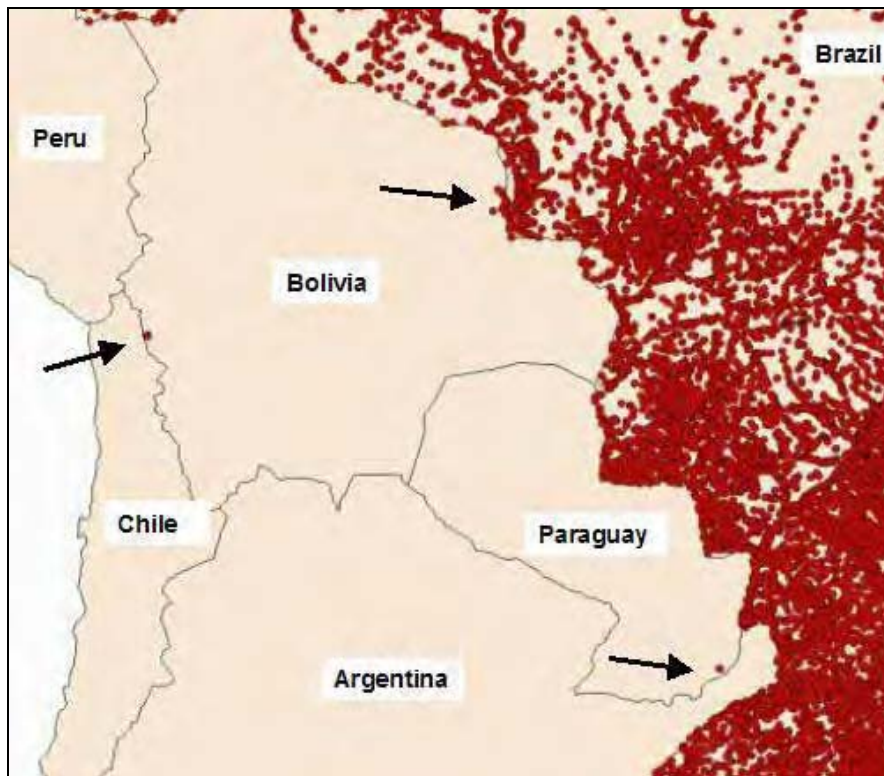


Fig. 15. Records from a Gazetteer of Brazilian place names showing a number of errors (arrowed), with one obvious error sitting on the Chile-Bolivian border and another in southern Paraguay.

A number of the tools, for example Diva-GIS (Hijmans *et al.* 2005) and the CRIA Data Cleaning tool (CRIA 2005) have routines that assist in identifying such errors.

The GIS can also be used to check that records fall outside a particular vegetation type, soil type or geology, etc. Some species are highly specific to certain geological types - limestone, sandstone, serpentinite (figure 16), for example. If you have the boundaries of these, any record that falls outside may be regarded as a possible outlier and flagged for further checking (Chapman *et al.* 2001). In figure 16, a species that only occurs on highly mineralised Serpentine soils is mapped and two records (marked 'a' and 'b') show up as likely errors. On checking, record 'a' only has the locality 'Goomeri' – the nearest town to the Serpentine outcrop, and has been geocoded with the latitude and longitude of the town. Record 'b' is quite near the outcrop and is likely misplaced due to the precision of the geocode given.

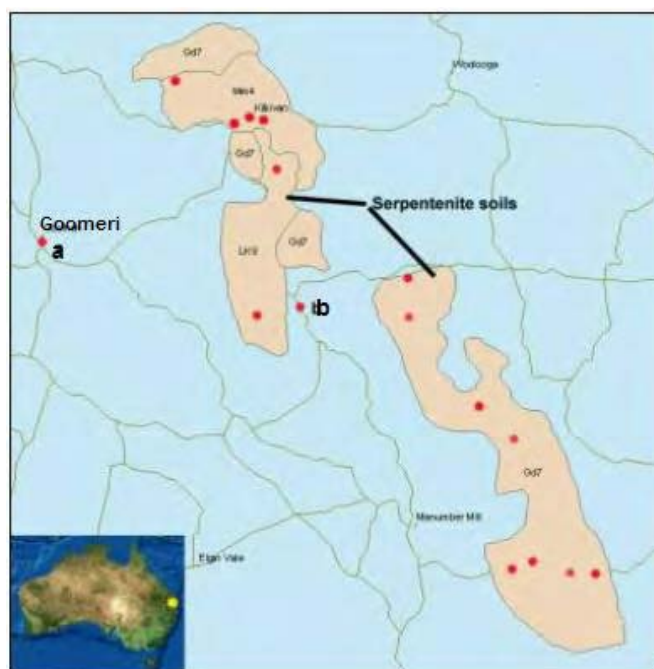


Fig. 16. Records of a species (red) that is only found on highly mineralised Serpentine soils. Records marked 'a' and 'b' have likely errors in geocoding.

The identification of collectors' itineraries (Chapman 1988, Peterson *et al.* 2003), allows for checking for possible error if, for example, the date of collection doesn't fit the particular pattern of that collector. This could be particularly useful for collectors from the 18th and 19th centuries prior to collectors being able to cover vast distances within the one day using helicopters, planes or motor vehicles. In the example in figure 17, collections between 22 and 25 February and in the first half of March should be in the Pentland-Lolworth area (circled), if outside that, they are likely to include errors in the date of the collection, or in the geocode (Chapman 1988). Again, using a GIS to map both the itinerary and the species' records can be very valuable. Another example is the use of animated GIS in Nepal to trace the routes of collectors along rivers (Lampe and Reide 2002).

Other uses of a GIS include for example, buffering of likely locations – e.g. streams for fish and aquatic plants, the coast for littoral species, altitudinal ranges for alpine species or others known to have a distinct altitudinal range. In this way, anything outside the buffer may need to be checked. Care needs to be exercised, as with the fish, for example, it may mean that those records outside the buffer zone are not errors at all, but the species may be occurring in small streams too small for mapping. These tests can generally only flag suspect records, and then it is up to individual checks to determine what may be real errors in the record, and what may be true outliers.

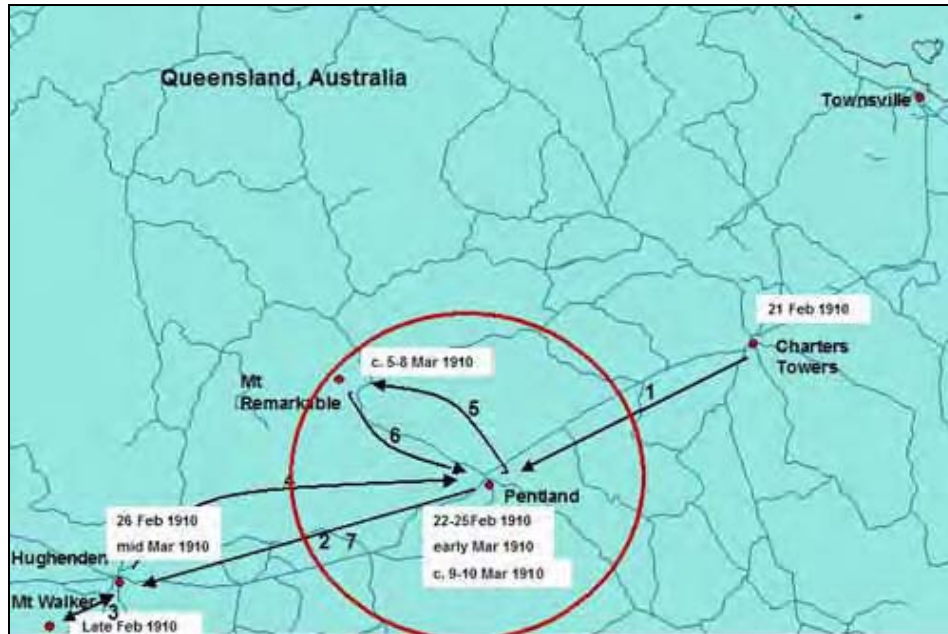


Fig. 17. *Collecting localities of Karl Domin in Queensland, Australia in 1910 (Chapman 1988). He travelled by train from Townsville to Hughenden, stopping at Charters Towers and Pentland. He then returned and spent about 10 days in the Pentland, Mount Remarkable, Lolworth area on horseback, before returning to Hughenden by train. Dates are only approximate.*

Outliers in geographic and environmental space

There are a number of methods for detecting outliers in data and these are outlined below. Natural history data is very diverse and generally does not conform to standard statistical distributions, and thus, as suggested by Maletic and Marcus (2000), more than one method is often necessary to capture most of the outliers.

Geographic Outlier Detection

A program from CRIA in Brazil (spOutlier) allows a user to type or cut and paste records into a box on the internet, link to a file, or submit an XML file of specimen records and receive information on geographic outliers. Records are submitted in the form: “id, latitude, longitude, altitude” and the program returns information on likely errors, both in textual form and on a map interface (Marino *et al.* in prep). It also allows the user to identify their data set as either an on-shore (terrestrial) or off-shore (marine) and again the program will return a list of mismatches. This is a unique program, and one that will prove very useful to biologists. It is also possible for users to submit a document on-line and have it returned, annotated with information on possible errors. An on-line version can be seen at <http://splink.cria.org.br/tools/> (CRIA 2004b).

In figure 14, the list of localities have returned four records with possible errors, 3 with possible errors in latitude, one with a possible error in longitude and one with a possible error in altitude. These points are then shown on the associated map with the records with possible errors identified in red.

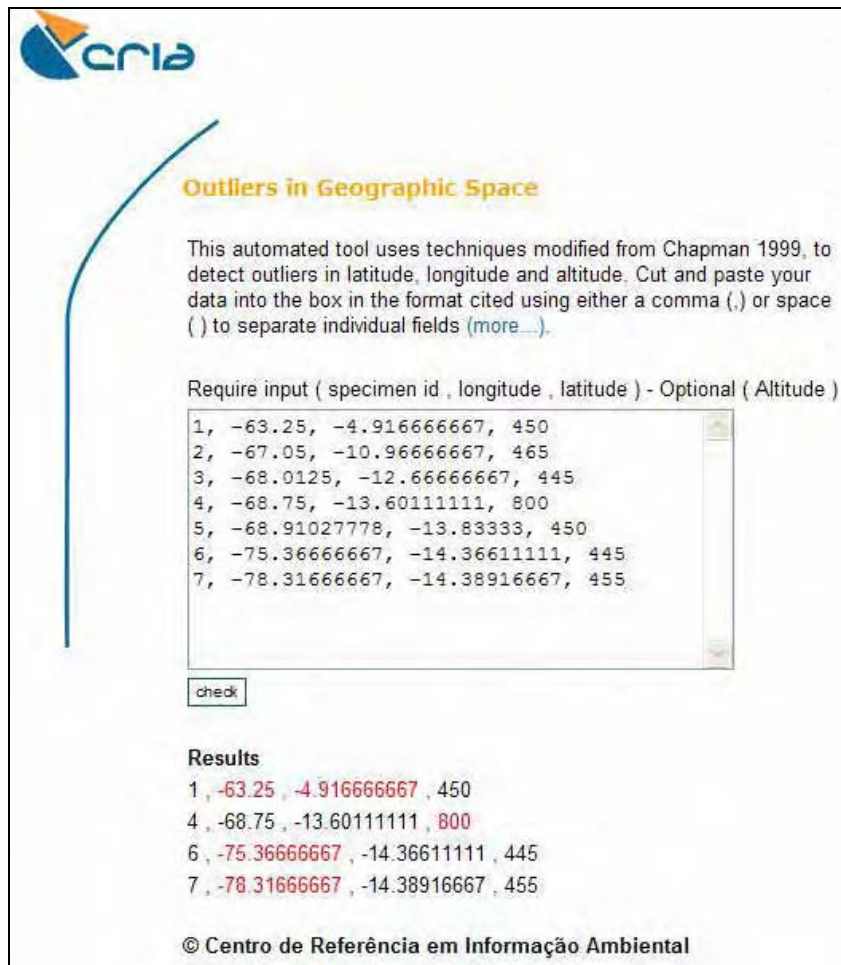


Fig. 18. Shows the prototype *Outliers in Geographic Space* system at CRIA identifying records 1, 4, 6 and 7 as having possible errors in geocoding.



Fig. 19. Map output associated showing identified suspect records (in red) from figure 14.

Publicly available programs using this method:

- **spOutlier-CRIA** (CRIA 2004b, Marino *et al.* in prep).
- **Data Cleaning-CRIA** (CRIA 2005).
- **Diva-GIS** (Hijmans *et al.* 2005)

Cumulative Frequency Curves

Early versions of the program BIOCLIM (Nix 1986, Busby 1991) were used to detect possible outliers by excluding records that fall outside the 90 percentile range of any element of the climate profile for the taxon, or by using cumulative frequency curves (Busby 1991, Lindemeyer *et al.* 1991) where the percentile figure can be varied. Although these techniques are still in use and are easy to use (Houlder *et al.* 2000, Hijmans *et al.* 2005) they do not allow for taxa that may not include any genuine errors, or that include many errors. They are also suspect for very small sample sizes (Chapman and Busby 1994, Chapman 1999).

A recent modification of the Diva-GIS software (Hijmans *et al.* 2005) has lead to the inclusion of the Reverse Jackknifing methodology (Chapman 1999) discussed below, and this has been linked to the Cumulative Frequency Curve with records identified under that method highlighted on the Cumulative Frequency curve for each parameter.

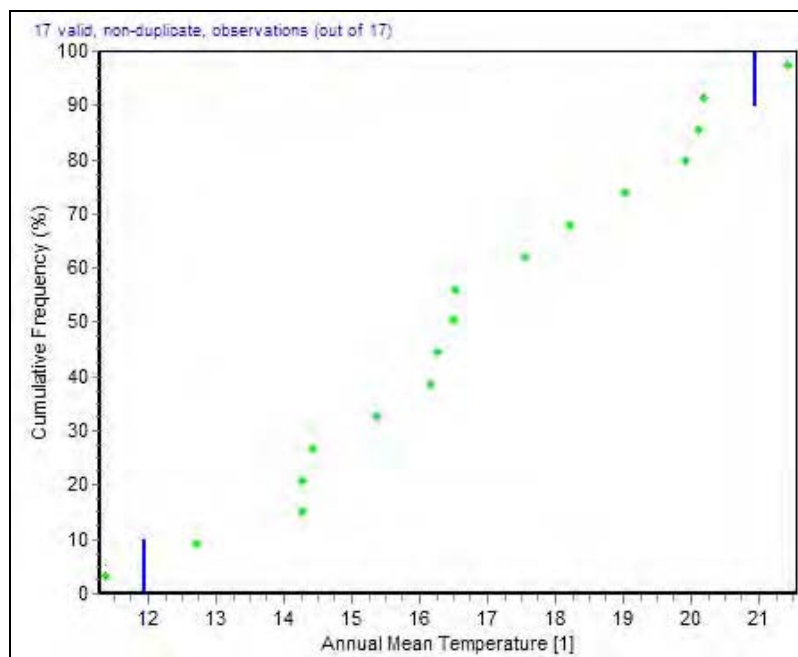


Fig. 20. Cumulative frequency curve used to detect outliers in climate space using Annual Mean Temperature. The Blue lines represent the 97.5 percentile, the point on the bottom left (or even the two to the bottom left), may be regarded as a possible outlier worth checking for error in the geocode.

Publicly available programs using this method:

- **Diva-GIS** (Hijmans *et al.* 2005)
- **ANUCLIM** (Houlder *et al.* 2000).

Principle Components Analysis

By using the scatter of points in a Principal Components Analysis of one climate layer against another one can identify possible outliers and thus possible errors in geocoding. It is a fairly powerful data validation method but unless the process is automated in some way to identify multiple outlier records the method can be quite tedious as one has to flick through however many combinations of climate components one is using.

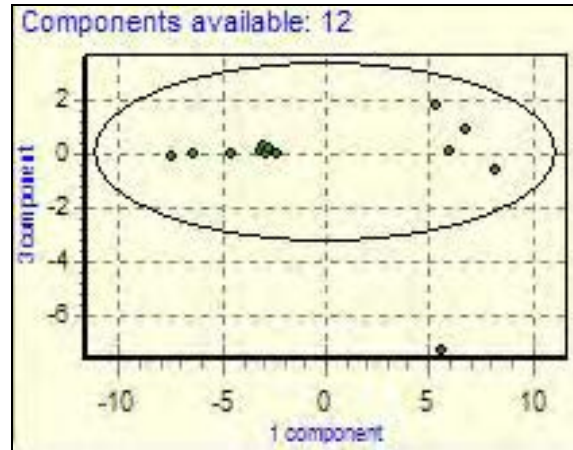


Fig. 21. *Principal Components Analysis, showing one point (in red) identified as an outlier and thus a possible error (from FloraMap, Jones and Gladkov 2001).*

Publicly available programs using this method:

- **FloraMap** (Jones and Gladkov 2001)
- **PATN** vers. 3.01 (Belbin 2004)

Cluster Analysis

The identification of outliers using clustering based on Euclidian or other distance measures can sometimes identify outliers that are not identified by methods at the field level (Johnson and Wichern 1998, Maletic and Marcus 2000). Cluster Analysis can be used to help identify multiple groups of like populations (using climate space or some other criteria), and can thus also be used to identify clusters that are isolated as either unicates or small groups separated by a significant distance from other clusters. Again, it is quite a valuable and seemingly robust methodology, but can depend very much on the cluster method used and can be computationally complex).

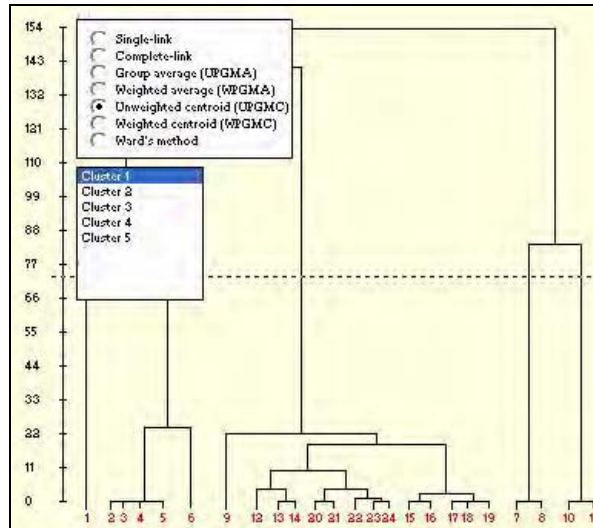


Fig. 22. Cluster Analysis showing a unicate cluster (#1 – in blue) which may be regarded as an outlier (from FloraMap, Jones and Gladkov 2001).

Publicly available programs using this method:

- **FloraMap** (Jones and Gladkov 2001)
- **PATN** Vers. 3.01 (Belbin 2004)

Climatic Envelope

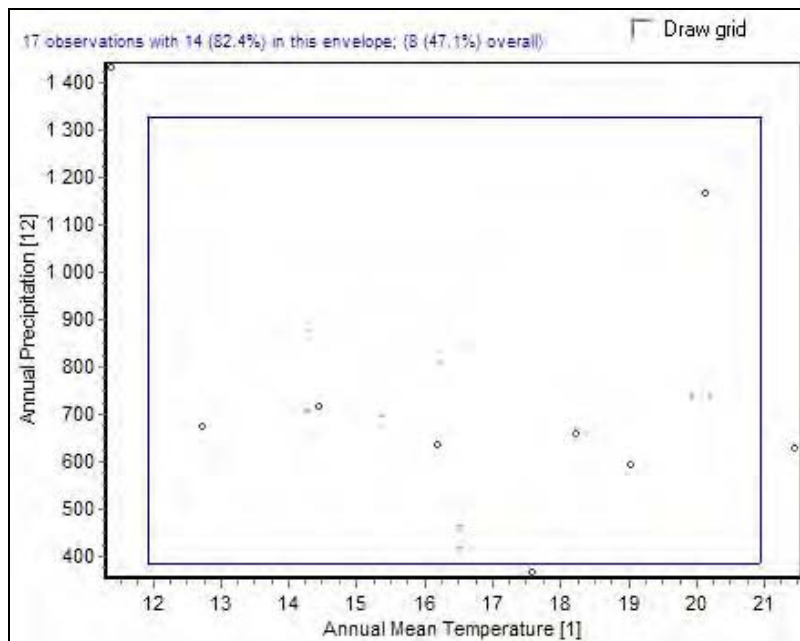


Fig. 23. Climatic envelope from BIOCLIM using a 97.5 percentile envelope for annual mean temperature and annual mean rainfall. Records marked in red are records that fall outside any one of the 64 possible envelopes.

The Climatic Envelope method is an extension of the cumulative frequency curve methodology mentioned above, but groups each of the climate layers into a multi-

dimensional box or envelope that can be examined two dimensions at a time, similar to the principal components analysis. Outliers in any of the cumulative frequency curves that make up the totality of climate layers can be identified in this manner.

Publicly available programs using this method:

- **Diva-GIS** (Hijmans *et al.* 2005)

Reverse Jackknife

This technique uses a modified reverse jackknifing to extract outliers at either end of an array of points in any one of a number of climate profiles. In 1992, the method was developed at ERIN in Australia to automatically detect outliers in climate space (Chapman 1992, 1999, Chapman and Busby 1994) and thus identify suspect records amongst the thousands of species being modelled at the time. The method has proved extremely reliable in automatically identifying suspect records, with a high proportion (around 90%) of those identified as being suspect, proving to be true errors.

$$\begin{array}{l}
 x < \bar{x} \\
 \text{if} \\
 y_{(i)} = (x_{(i+1)} - x_{(i)})(\bar{x} - x_{(i)}) \\
 \text{else} \\
 y_{(i)} = (x_{(i+1)} - x_{(i)})(x_{(i+1)} - \bar{x}) \\
 \text{then} \\
 C = \frac{y_{(i)}}{\sqrt{\frac{\sum_{i=1}^n (y_{(i)} - \bar{y})^2}{n-1}}}
 \end{array}$$

Fig. 24. Formula for determining the Critical Value (C) in an outlier detection algorithm where C = Critical Value (from Chapman 1999). This formula has been used in Australia since 1992 for detecting outliers in environmental (climate) space. The formula has recently been modified (2005) by dividing the value of C by the range of 'x' and has been incorporated into Diva-GIS version 5.0 (Hijmans *et al.* 2005). This has improved its reliability for use with criteria with large values such as rainfall, elevation, etc.

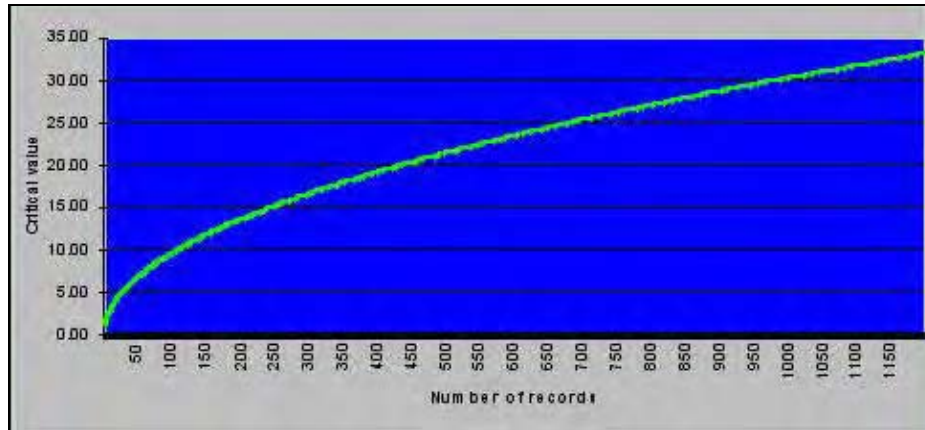


Fig. 25. Threshold Value Curve ($T=0.95(\sqrt{n})+0.2$ where 'n' is the number of records). Values above the curve are regarded as "suspect", values below the curve as "valid" (from Chapman 1999).

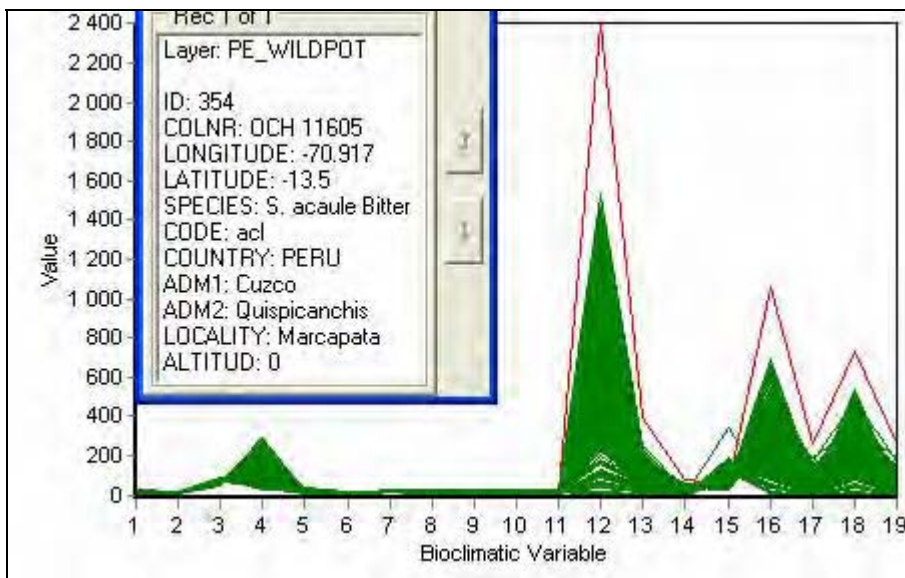


Fig. 26. Outlier Detection algorithm in Diva-GIS using Reverse Jackknifing. The program has identified one possible outlier (using the selected option to show only records that were outliers in at least 6 (of 19 possible) criteria).

Publicly available programs using this method:

- **Diva-GIS** (Hijmans *et al.* 2005)
- Also being programmed into the new BioGeomancer toolkit to be available mid 2006

Parameter Extremes

Parameter Extremes is a similar method to the Climatic Envelope method and identifies the record at the extremes of each Cumulative Frequency curve and bundles them into an output log file. In this way one can identify particular records that are extremes in more than one climate parameter.

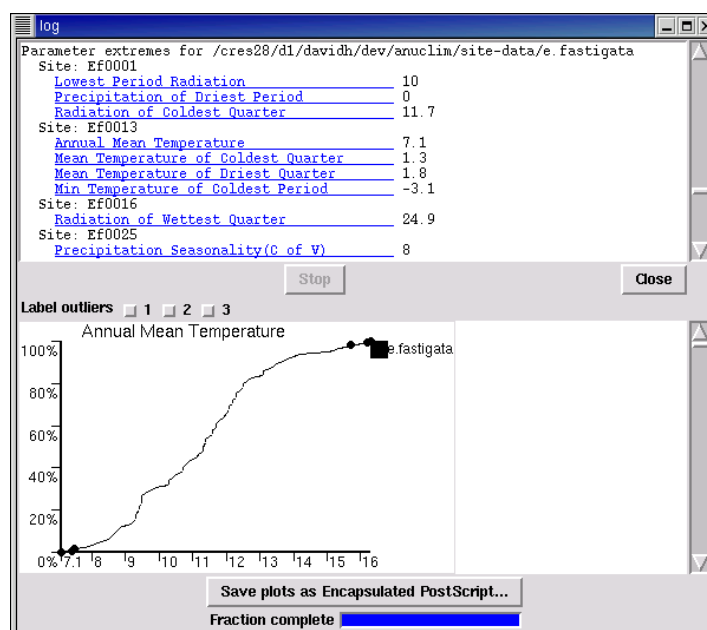


Fig 27. Log file of *Eucalyptus fastigata* from ANUCLIM Version 5.1 (Houlder *et al.* 2000) showing the parameter extremes (top) and one associated species accumulation curve (bottom).

Publicly available programs using this method:

- ANUCLIM (Houlder *et al.* 2000).

Other Methods

Many of the methodologies listed below are simple and are available in many standard statistical packages. Some do not seem to have been used for detecting errors in biological data, but examples with similar types of data indicate that they may be worth trying. A number of these and other methods are elaborated in Legendre and Legendre (1998). A number of other outlier detecting methods that may be worth trying, can be found in the publication by Barnett and Lewis (1994).

i. Standard Deviations from the Mean

Perhaps the most promising of these other methods would be to look at a varying number of standard deviations from the mean based on Chebyshev's theorem (Barnett and Lewis 1994). Maletic and Marcus (2000) tested a number of deviations from the mean using 5000 naval personnel records, with 78 fields of the same type (dates) and found that using 5 times the Standard Deviation generated the best results. Testing would need to be carried out on a number of collection datasets, and especially testing with much lower numbers of records than that used by Maletic and Marcus. Preliminary tests with low numbers by myself using elevation has so far not looked promising.

Deviations from the Median

Another group of non-parametric statistical tests use relationships with the median rather than the mean. Two possible methods are the Mann-Whitney U test and the Kuskall-Wallis test which look at the alternate hypothesis that two (Mann-Whitney), or three or more (Kuskall-

Wallis), populations differ only in respect to the median (Barnett and Lewis 1994, Lowry 2005). I have not seen examples of these applied to the detection of outliers in species-occurrence data, but they may be worth testing.

Use of Modelled Distributions

Distribution models derived from species distribution modelling such as those produced using GARP (Stockwell and Peters 1999, Pereira 2002) or Lifemapper (University of Kansas 2003b), could be used to identify new records that fall outside the predicted distribution. This method, although easy to use, is limited by the quality of the predicted distribution. If all records for a species had not been used to develop the model, then there may be deficiencies in that model. Also, using just the outer boundaries of the distribution does not take into account the scattered nature of good models that identify suitable niches within the broad totality of the geographic distribution.

Pattern Analysis

Pattern Analysis can be used to identify records that do not conform to existing patterns in the data. A variety of methods can be used for analysis of the patterns, including Association, Partitioning, Classification, Clustering, Ordination and use of Networks such as minimum spanning trees (Belbin 2004). Some of these methods have been discussed in more detail above. A pattern can generally be defined as a group of records that have similar characteristics (Maletic and Marcus 2000), but the choosing of the “right reference pattern” if such exists, can have an influence on the results (Weiher and Keddy 1999).

Publicly available programs using this method:

- **PATN** (Belbin 2004)

Ordinal Association Rules

Association rules attempt to find ordinal relationships that tend to hold over a large percentage of records (Marcus *et al.* 2001). They can be used for both categorical data and quantitative data. Simply put, they look for patterns such as if $A < B$ most of the time, then if $A > B$ in a record, then it is likely to be an error. With quantitative data, the rules can be used in conjunction with other statistical methods that use the mean, median, standard deviation and percentile ranges for outlier detection. With this method, the larger the number of records, the better the results, and in many cases could be used across whole databases rather than just within one species record. Uses could be such as, if species A occurs in Vegetation type B most of the time, then a record that has the information that it occurs in Vegetation type C may be an error. Or all records collected by a collector should not be within 15 years of the collector’s birth date, or greater than 100 years of their birth date, or later than their death date. Such rules could also be used in conjunction with a collector’s likely range (see above). For example, if a collection was before 1900, then two collections collected on the same day should not be greater than x kilometres apart.

Publicly available programs using this method:

- **PATN** (Belbin 2004).

Descriptive Data

Checking for errors in Descriptive Data is more difficult to cover here because of the quite diverse nature of what may be included in such databases. The structured nature of these databases, however, allow for more rule setting when the databases are set up.

Database design

The key to maintaining good data quality with descriptive databases is to follow good design procedures, and where possible design the databases following standards such as DELTA (Dallwitz *et al.* 1993) or the new SDD (Structure of Descriptive Data) standard (<http://160.45.63.11/Projects/TDWG-SDD/>) that is being developed by the Taxonomic Databases Working Group (TDWG).

Edit controls

Because of the structured nature of descriptive databases, they lend themselves to the use of edit controls. For example, most descriptive data fields have various constraints built in, and often have a well-developed set of characters from which the entries are chosen. Errors can still occur, however, especially with continuous data where units may be confused (e.g. millimetres and centimetres). Units used should be recorded, and preferably in a separate field as recommended in the SDD standard. Also – standardisation of units within one database should be carried out wherever possible – i.e. agree to use mm throughout, or cm, etc. rather than mix and match which can lead to errors, especially when entry of data is carried out by multiple operators. Tests can be carried out on these fields to look at extremes (e.g. by using cumulative frequency curves as described under the *Spatial Data* above), looking at outliers using Standard Deviations from the mean or median, etc. Often, by graphing the results one can also identify records that are possible errors. Some other error types that may be used to identify errors include (after English 1999).

- **Missing Data Values**
Searching for empty fields where values should occur. Where there is need for a “null” or missing value in a field it is good practice to record the reason for the null value in a separate field – for example “not relevant, not measured or unknown”..
- **Incorrect Data Values**
This involves searching for typographic errors, transposition of key strokes, data entered in the wrong place (e.g. alphanumeric characters entered into numerical fields), and data values forced into a field that requires a value, but for which the data entry operator doesn’t know the value so adds a dummy value. Dummy values are sometimes added into fields to “trick” statistical methods where empty fields or zero values are not allowed. This should be done with care.
- **Nonatomic Data Values**
Searching for fields where more than one fact is entered.
- **Domain Schizophrenia**
Searching for fields used for purposes for which they may not have been intended.
- **Duplicate Occurrences**
Searching for values that may refer to the same real world value. This can occur quite commonly when combining two databases that have used different terminologies.
- **Inconsistent Data Values**
Occurs where two related databases may not use the same values lists, and when combined show inconsistencies. This is where the use of transfer standards such as the SDD standard mentioned above come into play.

Documentation of Error

As mentioned in the associated document on *Principles of Data Quality* (Chapman 2005a), documentation of error and error checking is essential to maintain data quality and to avoid duplication of error checking. Without good documentation, users cannot determine the fitness of the data for use.

It is of very little use to anyone if checks of data quality are carried out, and corrections made, if they are not fully documented (Chapman 2005a). A data correction audit trail needs to be maintained as there is always the possibility that perceived errors are not errors at all, and that changes that are made, add new error. Without an audit trail, it may not be possible to undo those “corrections”. This is especially important where these checks are being carried out by other than the originator of the data (Chapman 2005a).

There are several ways of developing audit trails (i.e. recording changes made to the database over time as well as recording what data quality control checks have been carried out and when). Audit trails are important so that errors can be recovered, curators and data managers don't carry out checks that have already been carried out, and so that alterations and additions to the data are documented for legal and other purposes (for example, informing users who may have used the data knowing what changes have been made since they last accessed the data). One way of creating audit trails is through the application of a temporal database where a series of time stamps are added, for example a transaction time stamp period during which a fact should be stored in the database (Wikipedia⁶). Another method is to do periodic XML exports of the data of records that have changed, or portions of the data where changes have been made.

As mentioned in Chapman (2005a):

“data quality checks carried out on data by a user may identify a number of suspect records. These records may then be checked and found to be perfectly good records and genuine outliers. If this information is not documented in the record, further down the line, someone else may come along and carry out more data quality checks that again identify the same records as suspect.”

Also as mentioned in the associated document on *Principles of Data Quality* (Chapman 2005a):

One of the ways of making sure that error is fully documented is to include it in the early planning stages of database design and construction. Additional data quality/accuracy fields can then be incorporated. Fields such as geocode accuracy, source of information for the geocode and elevation, fields for who added the information – was the geocode added by the collector using a GPS, or a data entry operator at a later date using a map at a particular scale, was the elevation automatically generated from a DEM, if so, what was the source of the DEM, its date and scale, etc. All this information will be valuable in later determining whether the information is of value for a particular use or not, and the user of the data can then decide.

In addition, fields on data validation – the “who, when, how and what” of validation checks carried out should be added to the database to track and audit the validation, error checking

⁶ http://en.wikipedia.org/wiki/Temporal_database

and data cleaning carried out on the database. Ideally, these would also be added at the record level as suggested above.

Visualisation of Error

There is still a long way to go to develop good error visualisation methods for primary species data. The two requirements of visualisation are

- Visualisation for error checking and cleaning;
- Visualisation for presentation.

The second of these – visualisation for presentation was covered in the associated document on *Principles of Data Quality* (Chapman 2005a).

GIS is the most common method of visualising spatial error for use in checking. Just by mapping primary species data and overlaying it with topographic layers can assist in detecting errors. GIS systems range from simple on-line systems used mostly for on-line mapping and information presentation through to stand-alone systems that vary from the simple to the highly sophisticated.

Many institutions already use GIS for mapping, and these are easily adaptable for use in error checking. Other institutions, however, do not use GIS routinely and consider the purchase of a GIS system beyond their means, but there are free GIS programs available that are easy to learn and simple to use and that will adequately carry out most of the requirements of small collections institutions. At least one of these – Diva-GIS (Hijmans *et al.* 2005) – has been specifically designed for use by small museums and herbaria and includes several error detecting methods described in this document, as well as modelling and visualisation algorithms.

For non-spatial data, one can best visualise error through the use of spreadsheets and graphs. A simple graph of values will quickly identify records that don't fit the patterns. Simple graphs are easy to set up and populate from the database as a standard error checking method.

There is a growing tendency in the spatial community to use techniques such as Monte Carlo Analysis to produce estimates of the likely extent and importance of error (Flowerdew 1991). Monte Carlo analyses lend themselves well to visualisations, and are a good way of conveying error to users. Although some common software that includes Monte Carlo methods have become quite expensive (e.g. Canoco 4.5 for Windows⁷ and S-Plus⁸), free alternatives do exist, for example the PopTools add-in for Microsoft Excel (Hood 2005).

Visualising accuracy

As mentioned under *Georeferencing* above, point records of primary specimen records are not really points, but have an error figure associated with them. By mapping the point with its associated accuracy, the “footprint”, a good understanding of what the collection actually means, and its relationship to the real world, can be visualised.

This is one area of research that needs urgently pursuing with respect to primary species data – the development of techniques to visualize uncertainty and to show footprints of accuracy. Instead of a collection record being represented as a point of latitude and longitude there is a need to include the accuracy associated with the record and thus present the location as its footprint – a circle, an ellipse, a polygon or even a grid. GIS techniques, such as buffering,

⁷ <http://www.microcomputerpower.com/>

⁸ <http://www.insightful.com/products/splus/default.asp>

provide a good tool for developing footprints such as along rivers or roads. The Biogeomancer program is looking at some aspects of this, but is unlikely to develop a fully operational system in the time available.

Cited Tools

1. Software resources

ANUCLIM

Description: A bioclimatic modelling package containing a suite of programs, including the most recent version of BIOCLIM. The program includes a number of methods for identifying errors in the input specimen data.

Version: 5.1 (2004).

Custodian: Centre for Resource and Environmental Studies (CRES), Australian National University, Canberra, Australia.

Cost: \$AUD1000.

Reference: Houlder et al. 2000.

Download: <http://cres.anu.edu.au/outputs/software.php>

BioLink

Description: A software package designed to manage taxon-based information such as nomenclature, distribution, classification, ecology, morphology, illustrations, multimedia and literature.

Version: 2.1 (2005).

Custodian: Australian National Insect Collection, CSIRO, Canberra, Australia.

Cost: Free.

Reference: Shattuck and Fitzsimmons 2000.

Download: <http://www.biolink.csiro.au/>.

BIOTA

Description: A Biodiversity Data Management system for biodiversity and collections data. Its easy-to-use graphical interface harnesses the power of a fully relational database.

Version: 2.03 (2004).

Custodian: Robert K. Colwell, Connecticut, USA.

Cost: Demo Version Free: Full version: \$US200-600.

Reference: Collwell 2002.

Download: http://viceroy.eeb.uconn.edu/Biota2Pages/biota2_download.html

Biótica

Description: Designed to handle curatorial, nomenclatural, geographic, bibliographic and ecological data to assist in the capture and updating.

Version: 4.0 (2003).

Custodian: CONABIO, Mexico City, Mexico.

Cost: \$US290.

Reference: Conabio 2002.

Download: http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/distribu_v4.0.html.

BRAHMS

Description: A database software for botanical research and collection management. It provides support with the management of names, collection curation and taxonomic research.

Version: 5.58 (2005).

Custodian: University of Oxford, Oxford, UK.

Cost: Free.

Reference: University of Oxford 2004.

Download: <http://storage.plants.ox.ac.uk/brahms/defaultNS.html>.

Desktop GARP

Description: A software package for prediction and analysis of wild species distributions.
Version: 1.1.3 (2004)
Custodian: University of Kansas, Lawrence, Kansas, USA and Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.
Cost: Free.
Reference: Pereira 2002.
Download: <http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1>.

Diva-GIS

Description: A geographic information system developed for the analysis of biodiversity data. It includes several simple modelling tools and includes a number of data quality checking algorithms.
Version: 5.0 (2005).
Custodian: R.J. Hijmans *et al.*, University of California, Berkeley.
Cost: Free
Reference: Hijmans *et al.* 2005
Download: <http://www.diva-gis.org>

eGaz

Description: A program developed to assist museums and herbaria to identify and add geocodes to their specimen records.
Custodian: Australian National Insect Collection, CSIRO, Canberra, Australia.
Cost: Free
Reference: Shattuck 1997.
Download: <http://www.biolink.csiro.au/egaz.html>

FloraMap

Description: A software tool for predicting the distribution of plants and other organisms in the wild.
Version: 1.02 (2003).
Custodian: Centro Internacional de Agricultura Tropical (CIAT), Columbia.
Cost: \$US100.
Reference: Jones and Gladkov 2001.
Download: <http://www.floramap-ciat.org/ing/floramap101.htm>.

GeoLocate

Description: A georeferencing program to facilitate the task of assigning geographic coordinates to locality data associated with natural history collections.
Version: 2.0 (2003).
Custodian: Tulane Museum of Natural History, Belle Chasse, LA, USA.
Cost: Free.
Reference: Rios and Bart *n.dat*.
Order: <http://www.museum.tulane.edu/geolocate/order.aspx>.

PATN

Description: A comprehensive and versatile software package for extracting and displaying patterns in multivariate data.
Version: 3.01 (2004).
Custodian: Blatant Fabrications Pty Ltd (Lee Belbin)
Cost: \$US299.
Reference: Belbin 2004.
Download: <http://www.patn.com.au/>.

PopTools

Description: PopTools is a versatile add-in for PC versions of Microsoft Excel that facilitates analysis of matrix population models and simulation and stochastic processes.

Version: 2.6.6 (2005).

Custodian: Greg Hood, Albany, W.A., Australia.

Cost: Free

Reference: Hood 2005

Download: <http://www.cse.csiro.au/poptools/>.

Specify

Description: A collection management system for natural history museums and herbaria.

Version: 4.6 (2004).

Custodian: Biodiversity Research Center, The University of Kansas, Lawrence, Kansas, USA.

Cost: Free

Reference: University of Kansas 2003a

Download: <http://www.specifysoftware.org/Specify/specify/download>.

2. On-line resources

BioGeoMancer

Description: A georeferencing service for collectors, curators and users of natural history specimens.

Custodian: Peabody Museum of Natural History, Connecticut, USA.

Reference: Peabody Museum *n.dat*.

Location: <http://www.biogeomancer.org>

Notes: The BioGeomancer project has recently (2005) been expanded to become a worldwide collaboration of natural history museums with the aim of improving tools for georeferencing and data quality checking. The tools should be available for general use by mid 2006 both as stand-alone products as well as Web Services.

Data Cleaning (CRIA)

Description: An on-line data checking and error identification tool developed by CRIA to help curators of datasets made available via the speciesLink distributed information system to identify possible errors in their databases. Errors include both nomenclatural and geographic.

Custodian: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.

Location: <http://splink.cria.org.br/dc>

Notes: Some of the algorithms developed in this tool (especially the geographic tools) are being incorporated into the BioGeomancer toolkit as part of a worldwide collaborative project due for completion in mid 2006.

geoLoc

Description: A tool to assist biological collections in georeferencing their data.

Custodian: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.

Location: <http://splink.cria.org.br/geoloc?&setlang=en>

Georeferencing Calculator

Description: A java applet created to aid in the georeferencing of descriptive localities such as found in museum-based natural history collections.

Custodian: University of California, Berkeley, CA, USA.

Location: <http://manisnet.org/manis/gc.html>

Lifemapper

Description: Screensaver software that uses the Internet to retrieve records of plants and animals from natural history museums and uses modelling algorithms to predict distributions.
Custodian: Biodiversity Research Center, The University of Kansas, Lawrence, Kansas, USA.
Location: <http://www.lifemapper.org/>

spOutlier

Description: An automated tool used to detect outliers in latitude, longitude and altitude, and to identify errant on-shore or off-shore records in natural history collections data.
Custodian: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.
Location: <http://splink.cria.org.br/outlier?&setlang=en>

3. Standards and Guidelines

DELTA

Description: The DELTA format (DEscription Language for TAXonomy) is a flexible method for encoding taxonomic descriptions for computer processing.
Standard: Adopted by TDWG as a standard for data exchange.
Reference: Dallwitz *et al.* 1993.
Location: <http://biodiversity.uno.edu/delta/>

HISPID

Description: Herbarium Information Standards and Protocols for Interchange of Data.
Custodian: Committee of Heads of Australian Herbaria. Adopted as a TDWG Standard.
Reference: Conn 1996, 2000.
Location: <http://plantnet.rbg Syd.nsw.gov.au/Hispid4/>

MaNIS Georeferencing Guidelines

Description: Contains information about assigning geographic coordinates, and maximum error distances for those coordinates, to locality descriptions.
Custodian: University of California, Berkeley, CA, USA.
Location: <http://manisnet.org/manis/GeorefGuide.html>.

Manual de Procedimientos para Georreferenciar

Description: Manual developed by CONABIO in Mexico as guidelines for georeferencing natural history collections. In Spanish with an English abstract being prepared.
Reference: CONABIO 2005.
Location: Not yet available electronically.

MaPSTeDI Georeferencing Guidelines

Description: Guide to the specimen georeferencing process in the MaPSTeDI project.
Custodian: University of Colorado Regents, Denver, CO, USA.
Location: <http://mapstedi.colorado.edu/geocoding.html>

Plant Names in Botanical Databases

Description: The purpose of this standard is to specify how scientific names of plants may be organised in botanical databases.
Custodian: Taxonomic Databases Working Group (TDWG)..
Location: <http://www.tdwg.org/plants.html>

SDD

Description: The SDD subgroup of TDWG was established to develop an international XML-based standard for capturing and managing descriptive data for organisms.

Custodian: Taxonomic Databases Working Group (TDWG)..

Location: <http://160.45.63.11/Projects/TDWG-SDD/index.html>

TDWG Standards

Description: The Taxonomic Databases Working Group (TDWG) has been developing standards for use with biodiversity data for many years. Standards have been, and are being developed for a range of issues related to the storage, documentation, and distribution of species and species-occurrence data

Custodian: Taxonomic Databases Working Group (TDWG).

Location: <http://www.tdwg.org/standrds.html>
<http://www.tdwg.org/subgroops.html>

Conclusion

Errores ad sua principia referre, est refellere
To refer errors to their origin is to refute them.
(Ref. 3 Co. Inst. 15)

The information age has meant that collections' institutions have become an integral part of the environmental decision making process and politicians are increasingly seeking relevance and value in return for the resources that they put into those institutions. It is thus in the best interests of collections' institutions that they produce a quality product if they are to continue to be seen as a value-adding resource by those supplying the funding.

Best practice for databased information in museums and herbaria and institutions maintaining survey and observational information means making the data as accurate and possible, and using the most appropriate techniques and methodologies to ensure that the data are the best they can possibly be. To ensure that this is the case, it is essential that data entry errors are reduced to a minimum, and that on-going data cleaning and validation are integrated into day-to-day data and information management protocols.

There is no such thing as good quality data or bad quality data (Chapman 2005a). Data are data, and their use will determine their quality. Nevertheless, data providers need to ensure that the data are as free from error as it is possible to make them. No one test alone will ever be sufficient to identify all errors in a dataset, and thus it is important to use a combination of methods that best fit the circumstances of the organisation using them and the data contained therein. In addition collaboration between institutions, data providers, scientists and IT professionals as well as the users of the data is needed to improve data quality not only within individual collection institutions, but also across the totality of collections as combination takes place.

Perhaps the most important data management practice is good documentation. No matter what tests have been carried out on the data, they should be fully documented. Only in this way, can users of the data be truly informed as to their nature and likely accuracy.

In this period of increasing data and information exchange, the reputation of a collections' institution is likely to hinge on the quality and availability of its information (Redman 1996, Dalcin 2004), rather than on the quality of its scientists, as has been the case in the past. This is a fact of life, and the two can no longer be separated. Good data and information management must run side by side with good science and together they should lead to good data and information.

References:

- Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* **5(1)**: 80-88.
- ABRS. 2004. *Australian Faunal Directory*. Canberra: Australian Biological Resources Study.
<http://www.deh.gov.au/biodiversity/abrs/online-resources/abif/fauna/afd/index.html> [Accessed 12 Apr. 2005].
- ANBG. 2003. *Australian Plant Name Index*. Canberra: Australian National Botanic Gardens.
<http://www.anbg.gov.au/apni/index.html> [Accessed 12 Apr. 2005].
- Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data*. Chichester, UK: Wiley and Sons.
- Beaman, R.S. 2002. Automated georeferencing web services for natural history collections **in** *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002*
<http://www.cria.org.br/eventos/tdbi/flora/reed> [Accessed 12 Apr. 2005]
- Beaman, R., Wiczorek, J. and Blum, S. 2004. Determining Space from Place for Natural History Collections in a Distributed Library Environment. *D-Lib Magazine* Vol. 10(5).
<http://www.dlib.org/dlib/may04/beaman/05beaman.html> [Accessed 12 Apr. 2005].
- Belbin, L. 2004. *PATN vers. 3.01*. Blatant Fabrications <http://www.patn.com.au> [Accessed 12 Apr. 2005].
- Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283-309.
- Berendsohn, W., Güntsch, A. and Röpert, D. (2003). *Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections*. Copenhagen, Denmark: Global Biodiversity Information Facility.
http://circa.gbif.net/Members/irc/gbif/digit/library?l=/digitization_collections/contract_2003_report/ [Accessed 13 Apr. 2005].
- BioCASE. 2003. *Biological Collection Access Service for Europe*. <http://www.biocase.org> [Accessed 12 Apr. 2005].
- Birds Australia. 2004. *Birds Australia Rarities Committee (BARC)*.
<http://users.bigpond.net.au/palliser/barc/barc-home.html> [Accessed 12 Apr. 2005].
- Bisby, F.A. 1994. *Plant Names in Botanical databases*. TDWG Standard. <http://www.tdwg.org/plants.html> [Accessed 12 Apr. 2005].
- Bisby, F.A., Zarucchi, J.L., Schrire, B.L., Roskov, Y.R., Heald, J. and White, R.J. (eds). 2002. *ILDIS World Database of Legumes* ver. 6.05. <http://www.ildis.org/> [Accessed 12 Apr. 2005].
- Blakers, M., Davies, S.J.J.F. and Reilly, P.N. 1984. *The Atlas of Australian Birds*. Melbourne: Melbourne University Press.
- Blum, S. 2001. *Georeferencing Natural History Collection Localities at the California Academy of Sciences*.
<http://www.calacademy.org/research/informatics/GeoRef/index.html> [Accessed 12 Apr. 2005].
- Brickell, C.D., Baum, B.R., Hettterscheid, W.L.A., Leslie, A.C., McNeill, J., Trehane, P., Vrugtman, F. and Wiersema, J.H. (eds) 2004. *International Code for Cultivated Plants* ed. 7. Edinburgh, U.K.: ISHS.
<http://www.actahort.org/books/647/> [Accessed 11 Apr. 2005].
- Brummitt, R.K. and Powell, C.E. (eds). 1992 *Authors of Plant Names*. Kew: Royal Botanic Gardens, Kew.
<http://www.ipni.org/index.html> [Accessed 12 Apr. 2005]
- Burbidge, A.A. 1991. Cost Constraints on Surveys for Nature Conservation **in** Margules, C.R. and Austin, M.P. (eds). *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Canberra: CSIRO.
- Burrough, P.A. and McDonnell, R.A. 1998. *Principals of Geographical Information Systems*. Oxford, UK: Oxford University Press.
- Busby, J.R. 1991. BIOCLIM – a bioclimatic analysis and prediction system. Pp. 4-68 **in** Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO.
- Chapman, A.D. 1988. Karl Domin in Australia **in** *Botanical History Symposium. Development of Systematic Botany in Australasia. Ormond College, University of Melbourne. May 25-27, 1988*. Melbourne: Australian Systematic Botany Society, Inc.
- Chapman, A.D. 1991. Australian Plant Name Index pp. 1-3053. *Australian Flora and Fauna Series* Nos 12-15. Canberra: AGPS.

- Chapman, A.D. 1992. Quality Control and Validation of Environmental Resource Data **in** *Data Quality and Standards: Proceedings of a Seminar Organised by the Commonwealth Land Information Forum, Canberra, 5 December 1991*. Canberra: Commonwealth Land Information Forum.
- Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 **in** Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.
- Chapman, A.D. *et al.* 2002. *Guidelines on Biological Nomenclature*. Canberra: Environment Australia. <http://www.deh.gov.au/erin/documentation/nomenclature.html> [Accessed 12 Apr. 2005].
- Chapman, A.D. 2004. Guidelines on Biological Nomenclature. Brazil edition. Appendix J to *Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota*. FAPESP/Biota process no. 2001/02175-5 March 2003 – March 2004. Campinas, Brazil: CRIA 11 pp. http://smlink.cria.org.br/docs/appendix_j.pdf [Accessed 12 Apr. 2005].
- Chapman, A.D. 2005a. *Principles of Data Quality*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/data_quality [Accessed 1 Aug. 2005].
- Chapman, A.D. 2005b. *Uses of Primary Species-Occurrence Data*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/uses_of_data [Accessed 1 Aug. 2005].
- Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 **in** Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.
- Chapman, A.D. and Milne, D.J. 1998. *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia
- Chapman, A.D., Bennett, S., Bossard, K., Rosling, T., Tranter, J. and Kaye, P. 2001. Environment Protection and Biodiversity Conservation Act, 1999 – Information System. *Proceedings of the 17th Annual Meeting of the Taxonomic Databases Working Group, Sydney, Australia 9-11 November 2001*. Powerpoint: http://www.tdwg.org/2001meet/ArthurChapman_files/frame.htm [Accessed 12 Apr. 2005].
- CHAH 2002. *AVH - Australian's Virtual Herbarium*. Australia: Council of Heads of Australian Herbaria. <http://www.chah.gov.au/avh/avh.html> [Accessed 12 Apr. 2005].
- Chrisman, N.R., 1991. The Error Component in Spatial Data. pp. 165-174 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
- Christidis, L. & Boles, W.E. 1994. *Taxonomy and Species of Birds of Australia and its Territories*. Royal Australasian Ornithologists Union, Melbourne. 112 pp.
- Clarke, K.C. 2002. *Getting Started with Geographic Information Systems*, 4th edn. Upper Saddle River, NJ, USA: Prentice Hall. 352 pp.
- Colwell, R.K. 2002. *Biota: The Biodiversity Database Manager*. Connecticut, USA: University of Connecticut <http://viceroy.eeb.uconn.edu/Biota> [Accessed 12 Apr. 2005].
- CONABIO. 2002. *The Biótica Information system*. Mexico City: Comisión nacional para el conocimiento y uso de la biodiversidad. http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/acerca_biotica.html [Accessed 12 Apr. 2005]
- CONABIO. 2005. *Manual de Procedimientos para Georreferenciar*. Mexico: Comisión para el Conocimiento y Uso de la Biodiversidad México (CONABIO).
- Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data*. Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens. <http://www.bgbm.org/TDWG/acc/hispid30draft.doc> [Accessed 10 Apr 2005]
- Conn, B.J. (ed.) 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data*. Version 4 – Internet only version. Sydney: Royal Botanic Gardens. <http://plantnet.rbg Syd.nsw.gov.au/Hispid4/> [Accessed 30 Jul. 2003].
- Croft, J.R. (ed.) 1989. *HISPID – Herbarium Information Standards and Protocols for Interchange of Data*. Canberra: Australian National Botanic Gardens.
- CRIA. 2002. *speciesLink*. Campinas: Centro de Referência em Informação Ambiental. <http://smlink.cria.org.br/> [accessed 12 Apr. 2005]

- CRIA. 2004a. *GeoLoc-CRIA*. Campinas: Centro de Referência em Informação Ambiental. <http://splink.cria.org.br/tools/> [Accessed 12 Apr. 2005].
- CRIA. 2004b. *spOutlier-CRIA*. Centro de Referência em Informação Ambiental. <http://splink.cria.org.br/tools/> [accessed 1 Mar. 2005]
- CRIA (2005), *speciesLink. Dados e ferramentas. Data Cleaning*. Centro de Referência em Informação Ambiental. <http://splink.cria.org.br/dc/> [Accessed 12 Apr. 2005].
- Dalcin, E.C. 2004. *Data Quality Concepts and Techniques Applied to Taxonomic Databases*. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf [Accessed 7 Jan. 2005].
- Dallwitz, M.J., Paine, T.A. and Zurcher, E.J. (1993). *User's guide to the DELTA System: a general system for processing taxonomic descriptions*. 4th edn. <http://delta-intkey.com/> [Accessed 12 Apr. 2005].
- DEH. 2005a. *Threatened Species*. Canberra: Department of Environment and Heritage. <http://www.deh.gov.au/biodiversity/threatened/species/index.html> [Accessed 12 Apr. 2005].
- DEH. 2005b. *Species Profile and Threats Database*. Canberra : Department of Environment and Heritage. <http://www.deh.gov.au/cgi-bin/sprat/public/sprat.pl> [Accessed 7 Apr. 2005].
- Dorr, L.J. 1997. *Plant Collectors in Madagascar and the Comoro Islands*. Kew, UK: Royal Botanic Gardens, Kew.
- English, L.P. 1999. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., New York.
- ESRI. 2003. *ArcSDE: The GIS Gateway to Relational Databases*. <http://www.esri.com/software/arcgis/arcinfo/arcscde/overview.html> [Accessed 12 Apr. 2005]
- Farr, E. and Zijlstra, G. (eds). *n.dat. Index Nominum Genericorum (Plantarum)*. On-line version. <http://ravenel.si.edu/botany/ing/> [Accessed 21 Jul. 2004].
- Flowerdew, R., 1991. Spatial Data Integration. pp. 375-387 **in:** Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems Vol. 1, Principals*: Longman Scientific and Technical.
- Froese, R. and Bisby, F.A. (eds). 2004. *Catalogue of Life 2004*. Los Baños, Philippines: Species 2000. <http://www.sp2000.org/AnnualChecklist.html> [Accessed 7 Apr. 2005].
- Froese, R. and Pauly, D. 2004. *Fishbase*. Ver. 05/2004. The Philippines: World Fish Center. <http://www.fishbase.org/> [Accessed 10 Apr. 2005].
- Fundación Biodiversidad 2005. *Proyecto Anthos – Sistema de información sobre los plantas de España*. <http://www.programanthos.org/> [Accessed 8 Apr. 2005].
- Gatrell, A.C. 1991. Concepts of Space and Geographical Data. pp. 119-134 **in:** Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems Vol. 1, Principals*: Longman Scientific and Technical.
- GBIF. 2003a. *What is GBIF?* http://www.gbif.org/GBIF_org/what_is_gbif [Accessed 12 Apr. 2005].
- GBIF. 2003b. *GBIF Work Program 2004*. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/GBIF_org/wp/wp2004/GB7_20WP2004-v1.0-approved.pdf [Accessed 12 Apr. 2005].
- GBIF. 2004. *Data Portal*. Copenhagen: Global Biodiversity Information Facility. <http://www.gbif.net/portal/index.jsp>. [Accessed 12 Apr. 2005].
- Geographic Names Board. 2003. *Guidelines for Naming of Roads*. Sydney: Geographic Names Board of New South Wales. http://www.gnb.nsw.gov.au/newsroom/road_naming_guideline.pdf [Accessed 21 Jul. 2004].
- Greuter, W. *et al.* 1984-1989. *Med-Checklist: a critical inventory of vascular plants of the circum-mediterranean countries*. 4 Vols. Botanical Garden and Botanical Museum Berlin-Dahlem.
- Hepper, F.N. and Neate, F. 1971. *Plant Collectors in West Africa*. Utrecht, The Netherlands: Oosthoek's Uitgeversmaatschappij.
- Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrentes, I. and Rojas, E. 2005 *DIVA-GIS Version 5. A geographic information system for the analysis of biodiversity data*. <http://www.diva-gis.org> [Accessed 30 Jul. 2004].
- Hobern, D. and Saarenmaa, H. 2005. *GBIF Data Portal Strategy*. Draft Version 0.14. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/portal_strategy_1/ [Accessed 7 Apr. 2005].

- Hood, G.M. 2005. *PopTools version 2.6.6*. Canberra: CSIRO Sustainable Ecosystems. <http://www.cse.csiro.au/poptools> [Accessed 13 Apr. 2005].
- Houlder, D. Hutchinson, M.J., Nix, H.A. and McMahaon, J. 2000. *ANUCLIM 5.1 Users Guide*. Canberra: Cres, ANU. <http://cres.anu.edu.au/outputs/anuclim.php> [Accessed 12 Apr. 2005].
- IAPT. 1997. *Names in Current Use for Extant Plant Genera ver. 1.0*. on-line version. International Association for Plant Taxonomy. <http://www.bgbm.org/iapt/ncu/genera/Default.htm> [Accessed 12 Apr. 2005].
- ICSM. 2001. *Guidelines for the Consistent Use of Place Names*. Intergovernmental Committee on Survey and Mapping: Committee for Geographic Names in Australia. http://www.icsm.gov.au/icsm/cgna/consistent_pnames.pdf. [Accessed 12 Apr. 2005].
- Index Herbariorum. (1954-1988) *Index Herbariorum Part 2: Collectors*. Various compilers. Utrecht/Antwerp, The Hague/Boston
- Part 2(1): Collectors A-D (1954). *Regnum Vegetabile* vol. 2 (A-D),
- Part 2(2): Collectors E-H (1957). *Regnum Vegetabile* vol. 9 (E-H),
- Part 2(3): Collectors I-L (1972). *Regnum Vegetabile* vol. 86 (I-L),
- Part 2(4): Collectors M (1976). *Regnum Vegetabile* vol. 93 (M),
- Part 2(5): Collectors N-R (1983). *Regnum Vegetabile* vol. 189 (N-R),
- Part 2(6): Collectors S (1986). *Regnum Vegetabile* vol. 114 (S),
- Part 2(7): Collectors T-Z (1988). *Regnum Vegetabile* vol. 117 (T-Z).
- IPNI. 1999. *International Plant Names Index*. <http://www.ipni.org/index.html> [accessed 12 Apr. 2005].
- IOPI. 2003. *Global Plant Checklist*. International Organization for Plant Information (IOPI). <http://www.bgbm.fu-berlin.de/IOPI/GPC/> [Accessed 12 Apr. 2005].
- Johnson, R.A. and Wichern, D.W. 1998. *Applied Multivariate Statistical Analysis*. 4th edn. New York, NY: Prentice Hall.
- Jones P.G. and Gladkov, A. 2001. *Floramap Version 1.01*. Cali, Colombia: CIAT. <http://www.floramap-ciat.org/ing/floramap101.htm> [Accessed 12 Apr. 2005].
- Koch, I. 2003. *Coletores de plantas brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 12 Apr. 2005].
- Lampe, K.-H. and Riede, K. 2002. *Mapping the collectors: the georeferencing bottleneck*. Poster given to TDWG meeting, Indaiatuba, Brazil. <http://www.cria.org.br/eventos/tdbi/bis/Poster-200dpi.html> [accessed 12 Apr. 2005].
- Legendre P. and Legendre L. (1998): Numerical Ecology. *Developments in Environmental Modeling* 20, Second English Edition, Elsevier, Amsterdam, 853p. <http://www.bio.umontreal.ca/legendre/numecol.html> [Accessed 12 Apr. 2005].
- Lindemeyer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. and Tanton, M.T. 1991. The Conservation of Leadbeater's Possum, *Gymnobelidus leadbeateri* (McCoy): A Case Study of the Use of Bioclimatic Modelling. *J. Biogeog.* **18**: 371-383.
- Lowry, R. 2005. Concepts and Applications of Inferential Statistics. *VassarStats: Web Site for Statistical Computation*. <http://faculty.vassar.edu/lowry/webtext.html> [Accessed 8 Apr. 2005]
- Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis pp. 200-209 in *Proceedings of the Conference on Information Quality (IQ2000)*. Boston: Massachusetts Institute of Technology. <http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf> [Accessed 12 Apr. 2005].
- MaNIS. 2001. *The Mammal Networked Information System*. <http://manisnet.org/manis> [Accessed 12 Apr. 2005].
- Marcus, A., Maletic, J.I. and Lin, K.-I. 2001. Ordinal Association Rules for Error Identification in Data Sets pp. 589-591 in *Proceedings of the 10th ACM Conference on Information and Knowledge Management (ACM CIKM 2001)*. Atlanta, GA. <http://www.cs.wayne.edu/~amarcus/papers/cikm01.pdf> [Accessed 12 Apr. 2005].
- Margules, C.R. and Redhead, T.D. 1995. *BioRap. Guidelines for using the BioRap Methodology and Tools*. Canberra: CSIRO. 70pp.
- Marino, A., Pavarin, F., de Souza, S. and Chapman, A.D. in prep. *Simple on line tools for geocoding and validating biological data*. To be submitted.

- Neldner, V.J., Crossley, D.C. and Cofinas, M. 1995. Using Geographic Information Systems (GIS) to Determine the Adequacy of Sampling in Vegetation Surveys. *Biological Conservation* 73: 1-17.
- Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes in Longmore, R.C. (ed). Atlas of Australian elapid snakes. *Australian Flora and Fauna Series No. 7*: 4-15. Canberra: Australian Government Publishing Service.
- NMNH. 1993. *RapidMap. Geocoding locality descriptions associated with herbarium specimens*. U.S. National Museum of Natural History and Bernice P. Bishop Museum, Honolulu. <http://users.ca.astound.net/specht/rm/> [Accessed 12 Apr. 2005].
- Peabody Museum. *n.dat. BioGeoMancer*. <http://www.biogeomancer.org> [Accessed 12 Apr. 2005].
- Peterson, A.T., Navarro-Siguenza, A.G. and Benitez-Diaz, H. 1998. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* 140: 288-294.
- Peterson, A.T., Stockwell, D.R.B. and Kluza, D.A. 2002. Distributional Prediction Based on Ecological Niche Modelling of Primary Occurrence Data pp. 617-623 in Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Washington: Island Press.
- Peterson, A.T., Navarro-Siguenza, A.G. and Pereira, R.S. 2003. Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British Ornithologists' Club* 124: 143-151. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PNP_BBOC_2004.pdf. [Accessed 12 Apr. 2005].
- Pfeiffer, U., Poersch, T. and Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing and Management* 32(6): 667-679.
- Platnick, N.I. 2004. *The World Spider Catalog*. New York: The American Museum of Natural History. <http://research.amnh.org/entomology/spiders/catalog81-87/INTRO3.html> [Accessed 12 Apr. 2005].
- Podolsky, R. 1996. *Software Tools for the Management and Visualization of Biodiversity Data*. NY, USA: United Nations Development Project. <http://www3.undp.org/biod/bio.html> [Accessed 13 Apr. 2005].
- Pollock, J.J. and Zamora, A. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of ACM* 27(4): 358-368.
- Redman, T.C. 1996. *Data Quality for the Information Age*. Artech House Inc.
- Redman, T.C. 2001. *Data Quality: The Field Guide*. Boston, MA: Digital Press.
- Rios, N.E. and Bart, H.L. Jr. *n.dat. GEOLocate. Georeferencing Software. User's Manual*. Belle Chasse, LA, USA: Tulane Museum of Natural History. http://www.museum.tulane.edu/geolocate/support/manual_ver2_0.pdf [Accessed 12 Apr. 2005].
- Roughton, K.G. and Tyckoson, D.A. 1985. Browsing with sound: Sound-based codes and automated authority control. *Information Technology and Libraries* 4(2):130-136.
- Ruggiero, M. (ed.) 2001. *Integrated Taxonomic Information System*. <http://www.itis.usda.gov/> [Accessed 12 Apr. 2005].
- Pereira, R.S. 2002. *Desktop Garp*. Lawrence, Kansas: University of Kansas Center for Research. <http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1> [Accessed 13 Apr. 2005].
- Shattuck, S.O. 1997. eGaz, The Electronic Gazetteer. *ANIC News* 11: 9. <http://www.ento.csiro.au/biolink/egaz.html> [Accessed 12 Apr. 2005].
- Shattuck, S.O. and Fitzsimmons, N. 2000. *BioLink, The Biodiversity Information Management System*. Melbourne, Australia: CSIRO Publishing. <http://www.ento.csiro.au/biolink/software.html> [Accessed 12 Apr. 2005].
- Steenis-Kruseman, M.J. van 1950. Malaysian Plant Collectors and Collections. *Flora Malesiana* Vol. 1. Leiden, The Netherlands.
- Stockwell, D. and Peters, D. 1999. "The GARP modelling system: problems and solutions to automated spatial prediction." *International Journal of Geographical Information Science* 13(2): 143-158.
- University of Colorado Regents. 2003a. *mapstedi. Geocoding*. Denver: University of Colorado MaPSTeDI project. <http://mapstedi.colorado.edu/geocoding.html> [Accessed 12 Apr. 2005].
- University of Colorado Regents. 2003b. *GeoMuse*. Denver: University of Colorado MaPSTeDI project. <http://www.geomuse.org/mapstedi/client/start.jsp> [Accessed 12 Apr. 2005].
- University of Kansas. 2003a. *Specify*. Biological Collections Management. Lawrence, Kansas: University of Kansas <http://www.specifysoftware.org/Specify/> [Accessed 12 Apr. 2005].

- University of Kansas. 2003b. *LifeMapper*. Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. <http://www.lifemapper.org/> [Accessed 12 Apr. 2005]
- University of Oxford. 2004. *BRAHMS. Botanical Research and Herbarium Management System*. Oxford, UK: University of Oxford <http://storage.plants.ox.ac.uk/brahms/defaultNS.html> [Accessed 27 Jul 2004].
- Weber, W.A. 1995. Vernacular Names: Why Oh Why?. *Botanical Electrical News* No. 109. <http://www.ou.edu/cas/botany-micro/ben/ben109.html> [Accessed 7 Apr. 2005].
- Weiher, E. and Keddy, P. (eds). 1999. *Ecological Assembly Rules: Perspectives, Advances, Retreats*. Cambridge, UK: Cambridge University Press. 418 pp.
- Wieczorek, J. 2001a. *MaNIS: Georeferencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS <http://manisnet.org/manis/GeorefGuide.html> [Accessed 12 Apr. 2005].
- Wieczorek, J. 2001b. *MaNIS: Georeferencing Calculator*. Berkeley: University of California, Berkeley - MaNIS <http://manisnet.org/manis/gc.html> [Accessed 12 Apr. 2005].
- Wieczorek, J. and Beaman, R.S. 2002. Georeferencing: Collaboration and Automation in *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* <http://www.cria.org.br/eventos/tdbi/bis/georeferencing> [Accessed 12 Apr. 2005].
- Wieczorek, J., Guo, Q. and Hijmans, R.J. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18(8): 745-767.
- Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.
- Williams, P.H., Marguiles, C.R. and Hilbert, D.W. 2002. Data requirements and data. sources for biodiversity priority area selection. *J. Biosc.* **27(4)**: 327-338.

Index

- Accountability, 8
- Acknowledgements, 68
- Altitude, 44
- ANUCLIM*, 50, 56, 62
- Audit, 8, 60
- Australian Virtual Herbarium, 22, 31
- Authority files, 14
- BioCase*, 22
- BIOCLIM*, 50, 53
- BioGeomancer*, 64
- BioGeoMancer*, 3, 9, 10, 32, 33, 34
- BioLink*, 9, 37, 62
- Biota*, 9, 62
- Biótica*, 9, 62
- BRAHMS*, 9, 62
- Chebyshev's theorem, 57
- Classification Domain, 3
- Climate Envelope Method, 53
- Cluster Analysis, 52
- Collecting localities*, 47
- Collectors' itineraries, 46
- Collectors' localities, 44
- Conabio, 30
- CRIA, 48, 68
- CRIA Data Cleaning*, 64
- Cumulative Frequency Curve, 50, 59
- Data cleaning
 - Definition, 1
 - Methods, 9
 - Principles, 5
- Data Cleaning module (CRIA), 42, 49
- Data cleansing, 1
- Data entry, 12, 17, 19, 22, 24, 26, 29
- Data scrubbing, 1
- Data validation, 1
- Data values
 - Duplicate Occurrences, 16, 59
 - Inconsistent, 16, 59
 - Incorrect, 15, 59
 - Missing, 15, 59
 - Nonatomic, 15, 59
- Database design, 11, 13, 19, 29, 59, 60
- Database Merging, 4
- DELTA*, 59, 65
- Descriptive Data, 59
- Digital Elevation Model, 44
- Diva-GIS*, 38, 39, 40, 41, 46, 49, 50, 53, 55, 61, 63
- Documentation, 8, 60
- Documentation of Error, 60
- Domain Schizophrenia, 16, 59
- Duplication
 - minimising, 6
- Edit controls, 31, 59
- Education*, 5, 7
- eGaz*, 3, 9, 37, 38, 63
- ERIN, 68
- Error
 - Nomenclatural, 3
 - Prevention, 3, 5
 - Spatial, 3
 - Taxonomic, 3
 - Where, 2
- Error checking, 1, 12, 15, 18, 20, 22, 25, 27
 - using collectors' itineraries, 46
 - using collectors' localities, 44
 - using databases, 42, 44
 - using Digital Elevation Models, 44
 - using gazetteers, 44
 - using geology, 46
 - using GIS, 44
 - using soil type, 46
 - using vegetation, 46
- Error correction, 1
- Error detection, 1
 - Ordinal Association Rules, 58
 - Pattern Analysis, 57
- Error visualisation, 61
- Feedback, 7
- FloraMap*, 51, 52, 63
- GARP*, 57, 63
- Gazetteers, 32, 37
- GBIF Portal, 22, 31
- Geocode, 30
 - checking, 42
 - definition, 29
 - validation, 42
- Geocode assignment
 - using database, 31
- Geocoding software, 32
- Geographic Information Systems, 44, 46, 61

geoLoc, 3, 9, 10, 34, 35, 64
 GeoLocate, 36
 GEOLocate, 3, 9
GeoLocatep, 63
 Georeferencing, 29
 Definition, 29
 Retrospective, 29
Georeferencing Calculator, 3, 65
 Georeferencing guidelines, 30
 MaNIS, 9, 30, 65
 MaPSTeDI, 7, 9, 65
 MaPSTeDI, 30
 HISPID, 9, 11, 12, 13, 26, 65
 Information Management Chain, 1, 5
 Intraspecific rank, 19
 Kuskall-Wallis test, 57
Lifemapper, 57, 65
 MaNIS Georeferencing Guidelines, 65
 Mann-Whitney U test, 57
 MaPSTeDI Georeferencing Guidelines, 65
 Merging, 4
 Monte Carlo Analysis, 61
 Names
 author, 24
 cleaning, 42, 49
 collectors, 26
 colloquial, 17
 common, 17
 identification, 11
 scientific, 13
 spelling of, 13
 unpublished, 22
 Nomenclatural error, 3
 Non-parametric statistics, 57
 On-line resources, 64
 Ordinal Association Rules, 58
 Organizing data, 5
 Outlier detection
 Climatic Envelope, 53
 Cumulative Frequency Curve, 50
 Parameter extremes, 56
 Principal Components Analysis, 51
 Reverse jackknifing, 54
 Standard deviations from mean, 57, 59
 Standard deviations from median, 57
 Outliers, 47
 climatic space, 54
 environmental, 47
 geographic, 47
 Parameter extremes, 56
 Partnerships, 6
 PATN, 51, 52, 57, 58, 63
 Pattern Analysis, 57
 Performance measures, 6
 Phonix, 15, 18
 Planning, 5
 Plant Names in Botanical Databases, 66
 Point-in-polygon test, 45
 Principal Components Analysis, 51
 Principles of data cleaning, 5
 Prioritisation, 6
 RapidMap Geocoder, 34
 Resources
 Guidelines, 65
 On-line, 64
 Software, 62
 Standards, 65
 Retrospective georeferencing, 29
 Reverse jackknifing, 54
 Scale, 44
 Setting targets, 6
 Skeleton-Key, 15, 18
 Software resources, 62
 Soundex, 15, 18, 25
 Spatial accuracy, 30
 Spatial Data, 28
 Spatial error, 3
 speciesLink, 22, 26, 31, 35
 Specify, 9, 64
 spOutlier, 48, 49, 65
 Standard deviations from mean, 57, 59
 Standard deviations from median, 57
 Structure of Descriptive Data Standard, 59, 66
 Taxonomic Databases Working Group, 13, 59, 66
 Taxonomic error, 3
 TDWG, 13, 59
 Tools, 62
 Training, 5, 7
 Transparency, 8
 Visualisation
 use for data presentation, 61
 use for error checking, 61
 Visualising accuracy, 61
 Visualising error, 61

Chapter 5

Georeferencing

Introduction.....	1
DEFINITION.....	1
PRINCIPLES OF BEST PRACTICE	1
Background.....	2
Collecting And Recording Data In The Field	5
Beginning The Georeferencing Process.....	12
Georeferencing Legacy Data.....	20
Maintaining Data Quality.....	33
References	39
Appendix: Guidelines For Georeferencing Locality Types	43
Glossary.....	73
Index to Chapter 5.....	79

This Chapter is equivalent to:

Chapman, A.D. and J. Wiczorek (eds). 2006. *Guide to Best Practices for Georeferencing*.
Copenhagen: Global Biodiversity Information Facility. 84 pp. ISBN: 87-92020-00-3 (available as
a standalone PDF from <http://www.gbif.org>)

Introduction

One of the outputs from the [BioGeomancer](#) project is a document on best practice for georeferencing biological species (specimen and observational) data. Several projects ([MaNIS](#), [MapSteDI](#), [INRAM](#), [GEOLocate](#), [NatureServe](#), [CRIA](#), [ERIN](#), [CONABIO](#), etc.) have previously developed guidelines and tools for georeferencing, and these provide a good starting point for such a document.

The document provides guidelines to the world's best practice for georeferencing such data, but it is important that organisations and institutions then produce their own internal document that incorporates the practices outlined in this document into their own working environment.

The document presents examples of how to georeference a range of different location types, and provides information and examples on how to determine the extent and maximum uncertainty distance for locations based on the information provided.

1. Definition

“The term best practice generally refers to the best possible way of doing something; it is commonly used in the fields of business management, software engineering, and medicine, and increasingly in government. [...] The [qualified] term, ‘best current practice’, often represents the meaning in a more accurate way, showing the possibility for future developments of ‘better practice’.” ([Wikipedia: Best Practice](#)¹).

2. Principles of Best Practice

- **Accuracy** – a measure of how well the data represent true values. It is good practice to quote a percentage area or an uncertainty in meters, or to draw an uncertainty polygon.

With georeferencing – this is currently mostly an uncertainty radius, however uncertainty polygons are beginning to be used in some circumstances. Uncertainty probability surfaces are also under consideration.

- **Effectiveness** – the likelihood that a work program achieves its desired objectives.

With georeferencing – this is the percentage of records for which the latitude and longitude can be accurately identified through use of BioGeomancer or in some other way.

- **Efficiency** – the ratio of output to input.

With georeferencing – this is the amount of effort that is needed to produce an acceptable output. It also refers to the amount of input data the user has to obtain to produce an acceptable result (e.g., gazetteers, collectors' itineraries, etc.).

- **Reliability** – related to accuracy, and refers to the consistency with which results are produced.

With georeferencing – it refers to the repeatability with which a georeference can be produced by the user for the same locality.

¹ Wikipedia: Best Practice <http://en.wikipedia.org/wiki/Best_practice>

- **Accessibility** – how accessible are the results to the users, public, etc.

With georeferencing – this is the ease with which users, other institutions, etc., can access the information for a particular locality that has already been georeferenced.

- **Transparency** – an announcement of the procedures for collection, analysis, reporting and update.

With georeferencing – this refers to the quality of the metadata and methodology by which a georeference was obtained for a particular locality.

- **Timeliness** – relates to the frequency of data collection, its reporting and updates.

With georeferencing – it largely refers to how often gazetteers are updated, or when the records are georeferenced and made available to others.

- **Relevance** – the data collected should meet the needs of the user – i.e., should fulfill the principle of “fitness for use”.

With georeferencing – it refers to the format of the output (i.e., does it include good metadata on the above topics).

In addition, an effective best practices document should:

- Align the vision, mission, and strategic plans in an institution to policies and procedures and gain the support of sponsors and/or top management.
- Use a standard method of writing (writing format) to produce professional policies and procedures within the institution.
- Satisfy industry standards.
- Satisfy the scrutiny of management and external/internal auditors.

This list is by no means exhaustive, but does cover most of the elements in identifying best practice.

Background

A number of projects have been working for many years on the development of guidelines and tools for improving the georeferencing of primary biodiversity data. This document largely draws on those initiatives and attempts to bring the results of all this previous work into one comprehensive best practices document. Without this background work, such a document would not be possible. For link locations see under ‘Key Documents and Links’ at the end of this Chapter.

BioGeoMancer Classic

The original [BioGeoMancer Classic](#) was developed by Reed Beaman, now at Yale University. This tool provides a georeferencing service for collectors, curators and users of natural history specimens. [BioGeoMancer Classic](#) can parse English language place name descriptions and provide a set of latitude/longitude coordinates associated with that description. It provides offset calculations for when a collection is georeferenced a given distance and cardinal direction from the nearest named place. For more details on how it works – see “What it does ...²”.

² BioGeoMancer Classic – What it does ... <<http://130.132.27.130/you/bgm-docs/what-it-does.html>>

MaNIS

With support from the National Science Foundation, seventeen North American institutions and their collaborators developed the [Mammal Networked Information System](#). The original objectives of [MaNIS](#) were to 1) facilitate open access to combined specimen data from a web browser, 2) enhance the value of specimen collections, 3) conserve curatorial resources, and 4) use a design paradigm that could be easily adopted by other disciplines with similar needs.

The MaNIS network has developed a number of [tools and guidelines](#) for assisting the georeferencing of collections in the MaNIS network. These documents and tools have been heavily drawn upon in this document.

MapSTeDI

The Mountains and Plains Spatio-Temporal Database Informatics Initiative ([MaPSTeDI](#)) was a collaborative effort between the University of Colorado Museum, Denver Museum of Nature and Science, and Denver Botanic Gardens to convert their separate collections into one distributed biodiversity database and research toolkit for the southern and central Rockies and adjacent plains. Unlike MaNIS or other projects, which have strong taxonomic focus and a distributed database federation outcome, MaPSTeDI had a regional focus and a distributed GIS mapping system outcome. Like other projects listed here, georeferencing was the essential first step in MaPSTeDI, providing the data that will be eventually analyzed spatially and temporally on the MaPSTeDI online GIS. The MaPSTeDI project also developed detailed [guidelines and tools](#) such as the [MaPSTeDI Georeferencing Protocols](#) and [Guide to Georeferencing](#), and these have been heavily relied upon in this document.

INRAM

The [Institute of Resource Analysis and Management](#) (INRAM) sought to increase the value of New Mexico museum specimen data by supporting an effort to georeference New Mexico specimen localities. Data that are georeferenced haphazardly are of little use to science, so the first goal of the INRAM Georeferencing Team was to develop a detailed, comprehensive protocol describing how best to determine the coordinates and uncertainty estimate to apply to a given locality. The INRAM team started by evaluating the protocol used by the [Mammal Networked Information System \(MaNIS\)](#) in which the Museum of Southwestern Biology (MSB) mammal division was participating, and determined that there were many ways it could be improved. In particular, INRAM created a more detailed list of locality types with a specific rule set for each as to how to determine coordinates and uncertainty. INRAM also sought to maximize the efficiency and accuracy of the georeferencing process. With help from the New Mexico Natural Heritage Program and the Museum of Southwestern Biology, INRAM developed a combined GIS and database system that made implementing the protocol much easier for the students doing the work. Together, the INRAM protocol and georeferencing software system allowed a semi-automated georeferencing process which provided accurate, rapid data capture and which left a detailed record of the methods and assumptions used to georeference each specimen.

GEOLocate

In March of 1995, Dr. Henry L. Bart received funding from the U.S. National Science Foundation to computerize and georeference the Tulane University Museum of Natural History Fish Collection. Georeferencing was accomplished by manually plotting each locality description on hardcopy USGS topographic maps and using a digitizing tablet to

register the maps and determine coordinates. Where possible, hand-plotted, hardcopy maps were compared to electronic versions of the same maps (USGS digital line graphs), allowing the technician to use a mouse to electronically capture the coordinates. Using this method, 15,000 locality descriptions for nearly 7 million specimens were georeferenced by one technician over a period of 18 months.

In February of 2002, Dr. Bart and Nelson Rios received funding from the the U.S. National Science foundation to develop a software package to facilitate georeferencing of natural history collections data, using the Tulane Fish Collection as a testbed. The result was [GEOLocate](#), a tool for comprehensive automated georeferencing of North American locality descriptions. Ongoing development involves expanding coverage to the entire world, multi-lingual support, user-defined pattern recognition, and collaborative georeferencing. GEOLocate is also being developed as a webservice for integration into the current development of [BioGeomancer](#).

ERIN

The Environmental Resources Information Network (ERIN) was established in the Australian Department of the Environment in 1989 and began funding the databasing and georeferencing of Australia's museum and herbarium collections. It established methods for assisting georeferencing, including the linking of records to Digital Elevation Models to determine elevation, and sophisticated methods for data checking and validation by searching for outliers in environmental space using niche modeling techniques. These have recently been upgraded in conjunction with the Centro de Referência em Informação Ambiental (CRIA) and Robert Hijmans, the author of the DIVA-GIS software.

Key Documents and Links

- Best Practices Guidelines for GPS Survey (NLWRA, Australia)
<http://www.nlwra.gov.au/toolkit/10/10-2.html>
- BioGeoMancer Classic
<http://classic.biogeomancer.org>
- Centro de Referência em Informação Ambiental (CRIA)
<http://www.cria.org.br>
- DIVA-GIS
<http://www.diva-gis.org>
- Environmental Resources Information Network (ERIN)
<http://www.deh.gov.au/erin/index.html>
- Examples of Good and Bad Localities
http://mvz.berkeley.edu/Locality_Field_Recording_examples.html
- GEOLocate – University of Tulane
<http://www.museum.tulane.edu/geolocate/>
- Institute of Resource Analysis and Management (INRAM)
<http://biodiversity.inram.org/>
- INRAM Protocol for Georeferencing Biological Museum Specimen Records
http://www.inram.org/modules/UpDownload/store_folder/Documents/INRAM_Biodiversity_Georeferencing_Project/Georeferencing_Guidelines_INRAM-V1.3_2004-03-01.pdf
- Mammal Networked Information System (MaNIS)
<http://manisnet.org/>

- MaNIS Documents
<http://manisnet.org/Documents.html>
- MaNIS/HerpNet/ORNIS Georeferencing Guidelines
[http://manisnet.org/manis GeorefGuide.html](http://manisnet.org/manis_GeorefGuide.html)
- Manual de Procedimientos para Georeferenciar, CONABIO, 2004. An internal georeferencing manual produced by the Comisión Nacional para el Conocimiento y Uso de la Biodiversidad ([CONABIO](http://www.conabio.gob.mx)), Mexico.
- The Mountains and Plains Spatio-Temporal Database Informatics Initiative - MaPSTeDI
<http://mapstedi.colorado.edu/index.html>
- MaPSTeDI Georeferencing Protocols
<http://mapstedi.colorado.edu/georeferencing-protocols.html>
- MaPSTeDI Guide to Georeferencing
<http://mapstedi.colorado.edu/georeferencing-howto.html>
- Museum of Vertebrate Zoology Informatics (MVZ) – University of California, Berkeley
<http://mvz.berkeley.edu/Informatics.html>
- MVZ Guide for Recording Localities in the Field:
http://mvz.berkeley.edu/Locality_Field_Recording_Notebooks.html
- Reasons Why it is Important to Take Good Locality Data (MVZ)
http://mvz.berkeley.edu/Locality_Field_Recording_important.html
- OGC Recommendations Document Pointer
<http://www.opengeospatial.org/specs/?page=recommendation>

Collecting and Recording Data in the Field³

Collecting data in the field sets the stage for good georeferencing procedures. Many new techniques now exist that can lead to quite accurately georeferenced locations; however it is important that the locations be recorded correctly in order to reduce the likelihood of error. We recommend that all new collecting events use a GPS for recording coordinates wherever possible, and that the GPS be set to a relevant datum (see below).

1. The Importance of Good Locality Data Recording

Good locality descriptions lead to more accurate georeferences with smaller uncertainty values and provide users with much more accurate and high quality data. When recording data in the field, whether from a map or when using a GPS, it is important to record locality information as well as the georeferences, so that later validation can take place if necessary.

One purpose behind a specific locality description is to allow the validation of coordinates, in which errors are otherwise difficult to detect. The extent to which validation can occur depends on how well the locality description and its spatial counterpart describe the same place. The highest quality locality description is one with as few sources of uncertainty as possible. By describing a place in terms of a distance along a path, or by two orthogonal distances from a place, one removes uncertainty due to imprecise headings. Choosing a reference point with small extent reduces the uncertainty due to the size of the reference

³ See also Museum of Vertebrate Zoology, Berkeley, California (2006) *MVZ Guide for Recording Localities in Field Notes* <http://mvz.berkeley.edu/Locality_Field_Recording_Notebooks.html>

point, and by choosing a nearby reference point, one reduces the potential for error in measuring the offset distances.

To make it easy to validate a locality, use reference points that are easy to find on maps or in gazetteers. At all costs, avoid using vague terms such as “near” and “center of” or providing only an offset without a distance such as “West of Albuquerque”.

In any locality that contains a named place that can be confused with another named place of a different type, specify the feature type in parentheses following the feature name.

Examples:

Locality example using distance and heading along a path:

E shore of Bolinas Lagoon, 3.1 mi NW via Hwy. 1 from intersection of Hwy. 1 and Calle del Arroyo in Stinson Beach (town), Marin Co., Calif.

Locality example using two cardinal offset distances from a reference point:

ice field below Cerro El Plomo, 0.5 km S and 0.2 km W of summit, Region Metropolitana, Chile.

2. Recording Localities

Provide a descriptive locality, even if you have geographic coordinates. The locality should be as specific, succinct, unambiguous, complete, and as accurate as possible, leaving no room for uncertainty in interpretation.

Localities used as reference points should be stable – i.e., places (towns, trig points, etc.) that will remain for a long time after the collection events. Do NOT use temporary locations or waypoints as the key reference location. You may have made an accurate GPS recording for the temporary location and then referenced future collections from that point (e.g., 200 m SE of the Land Rover), and that may make perfect sense for that series of collections. It is meaningless, however, when those collections are later broken up and placed in a museum under a taxonomic arrangement, and no longer have a link to where the ‘Landrover’ was.

If recording locations along a path (road, river, etc.) it is important to also record whether the distances were measured along the path (‘by road’) or as a direct line from the origin (‘by air’).

Hint: The most specific localities are those described by a) a distance and heading along a path from a nearby and well-defined intersection, or b) two cardinal offset distances from a single persistent nearby feature of small extent.

3. Recording Coordinates

Coordinates are a convenient way to define a locality that is not only more specific than is otherwise possible with a description, but that is also readily usable in GIS applications. Always include as many decimals of precision as given by the coordinate source. A measurement in decimal degrees given to five decimal places is more precise than a measurement in degrees minutes seconds to the nearest second, and more precise than a measurement in degrees decimal minutes given to three decimal places (see Table 4). Some new GPS receivers now provide for recording data in decimal seconds and this (to two decimal places) provides a precision comparable to that of decimal degrees.

Whenever practical, provide the coordinates of the location where collecting actually occurred (see *Extent*, below). If reading coordinates from a map, use the same coordinate system as the map. The datum is an essential part of a coordinate description; it provides the frame of reference. When using both maps and GPS in the field, set the GPS datum to be the

same as the map datum so that your GPS coordinates will match those on the map. Be sure to record the datum used.

Specific projects may require particular coordinate systems, but we find geographic coordinates in decimal degrees to be the most convenient system for georeferencing. Since this format relies on just two attributes, one for latitude and the other for longitude, it provides a succinct coordinate description with global applicability that is readily transformed to other coordinate systems as well as from one datum to another. By keeping the number of recorded attributes to a minimum, the chances for transcription errors are minimized (Wieczorek *et al.* 2004).

Hint: Decimal degrees are preferred when reading coordinates from a GPS, however see Note under *Using a GPS*, below.

Hint: If using UTM coordinates, always record the UTM Zone.

4. Using a GPS

GPS (Global Positioning System) technology uses triangulation to determine the location of a position on the earth's surface. The distance calculated is the range between the GPS receiver and the GPS Satellites (Van Sickle 1996). As the GPS satellites are at known locations in space, the position on earth can be calculated. A minimum of four GPS satellites is required to determine the location of a position on the earth's surface (McElroy *et al.* 1998, Van Sickle 1996). This is not generally a limitation today, as one can often receive seven or more satellites in most locations on earth, however, historically the number of satellites receivable was not always sufficient. Prior to May 2000, most GPS units used by civilians were subject to "Selective Availability". The removal of this signal degradation technique has greatly improved the accuracy that can generally be expected from GPS receivers (NOAA 2002).

To obtain the best possible accuracy, the GPS receiver must be located in an area that is free from overhead obstructions and reflective surfaces and have a good field of view to the horizon (for example, they do not work very well under a heavy forest canopy). The GPS receiver must be able to record signals from at least four GPS satellites in a suitable geometric arrangement. The best arrangement is to have "*one satellite directly overhead and the other three equally spaced around the horizon*" (McElroy *et al.* 1998). The GPS receiver must also be set to an appropriate datum for the area, and the datum used recorded (Chapman *et al.* 2005a).

GPS accuracy: Most GPS devices are able to report a theoretical horizontal accuracy based on local conditions at the time of reading. For highly specific localities, it may be possible for the potential error in the GPS reading to be on the same order of magnitude as the extent of the locality. In these cases, the GPS accuracy can make a non-trivial contribution to the overall uncertainty in the position given by the coordinates.

Prior to the removal of Selective Availability, the accuracy of *Hand-held GPS* receivers as used by most biologists and observers in the field was around 100 meters or worse (McElroy *et al.* 1998, Van Sickle, 1996, Leick 1995). Since then, however, the accuracy of GPS receivers has improved and today, most manufacturers of hand-held GPS units promise errors of less than 10 meters in open areas when using four or more satellites. The accuracy can be improved by averaging the results of multiple observations at a single location (McElroy *et al.* 1998), and some modern GPS receivers that include averaging algorithms can bring the accuracy down to around five meters or maybe even better. NOAA (2001) suggests that GPSs without differential (see below) may be as accurate as 10-15 meters, depending on the

receiver being used, satellite configuration and atmospheric conditions, but that this is at the better end of the scale.

The use of *Differential GPS* (DGPS) can improve the accuracy considerably. DGPS uses referencing to a GPS Base Station (usually a survey control point) at a known location to calibrate the receiving GPS. This works through the Base Station and hand-held GPS referencing the satellites' positions at the same time and thus reduces error due to atmospheric conditions. In this way, the hand-held GPS applies the appropriate corrections to the determined position. Depending on the quality of the receivers used, one can expect an accuracy of between 1 and 5 meters. This accuracy decreases as the distance of the receiver from the Base Station increases. Again, averaging can further improve on these values (McElroy *et al.* 1998). For example, the U.S. Coast Guard's DGPS has a stated horizontal accuracy of ± 10 meters (95%). In other words, 95 percent of the time a position determined using DGPS will be within 10 meters of its true position on the earth. Under certain conditions, mariners may observe better than 10-meter accuracy (NOAA 2001).

The Wide Area Augmentation System (WAAS) is a GPS-based navigation and landing system developed for precision guidance of aircraft (Federal Aviation Administration 2004). WAAS uses ground-based antennae with precisely known locations to provide greater positional accuracy for GPSs. Similar technologies such as Local Area Augmentation System (LAAS) are also being developed to provide even finer precision.

Even greater accuracies can be achieved using either *Real-time Differential GPS* (McElroy *et al.* 1998) or *Static GPS* (McElroy *et al.* 1998, Van Sickle 1996). *Static GPS* uses high precision instruments and specialist techniques and is generally employed only by surveyors. Surveys conducted in Australia using these techniques reported accuracies in the centimeter range. These techniques are unlikely to be extensively used with biological record collection due to the cost and general lack of requirement for such precision.

Note! Set your GPS to report locations in decimal degrees rather than make a conversion from another coordinate system as it is usually more precise, better and easier to store, and saves later transformations which may introduce error.

Note2! An alternative where reference to maps is important, and where the GPS receiver allows it, is to set the recorder to report in degrees, minutes, and decimal seconds.

5. Recording Datum

Except under special circumstances (the poles, for example), coordinates without a datum do not uniquely specify a location. Confusion about the datum can result in positional errors of hundreds of meters.

When using a GPS, it is important to set and record the Datum being used. See discussion below under *Calculating Uncertainties*.

Note! If you are not basing your locality description on a map, set your GPS to report coordinates using the WGS84 datum. Record that fact in all your documentation.

6. Recording Elevation

Supplement the locality description with elevation information if this can easily be obtained. It is preferable to use a barometric altimeter if available. Alternatively, obtain the elevation

from a Digital Elevation Model (usually done retrospectively in the laboratory), or by using the contours and spot height information from a suitable scale map of the area. Record the method used in Remarks.

Note! "Elevation markings can narrow down the area in which you place a point. More often than not, however, they seem to create inconsistency. While elevation should not be ignored, it is important to realize that elevation was often measured inaccurately and/or imprecisely, especially early in the 20th century. One of the best uses of elevation in a locality description is to pinpoint a location along a road or river in a topographically complex area, especially when the rest of the locality description is vague."

(MaPSTeDI 2004)

Under normal conditions, GPS devices are much less accurate for recording elevation than horizontal distances, and they do not report the altitudinal accuracy. It is important to note that the height displayed by a GPS receiver is actually the height in relation to an ellipsoid as a model of the Earth's surface, and not a height based on mean sea level, or to a standard height datum such as the Australian Height Datum. In Australia, for example, the difference between altitudes reported from a GPS receiver and mean sea level can vary from -35 to +80 meters and tends to vary in an unpredictable manner (Chapman *et al.* 2005, McElroy *et al.* 1998, Van Sickle 1996).

If elevation is a defining part of the locality description, be sure to use a reliable source for this measurement (barometric altimeter, trustworthy map, or Digital Elevation Model at suitable scale), and specify the source under references. It is not recommended that elevation be determined using a GPS.

Hint: A barometric altimeter, when properly calibrated, is much more reliable than a GPS for obtaining accurate elevations. It is not recommended that elevation be determined using a GPS. See remarks above under **Using a GPS** about the error inherent in using a GPS to determine elevations.

7. Recording Headings

It is important when using a compass to record headings, that adjustments be made to record True North and not Magnetic North. The differences between True North and Magnetic North vary in different parts of the world, and in some places can vary greatly across a very small distance. The differences also change over time. For example, in an area about 250 km NW of Minneapolis in the United States, the anomolous declination changes from 16.6° E to 12.0° W across a distance of just 6 km (Goulet 2001).

The National Geophysical Data Center (NGDC) in the USA has an on-line calculator⁴ that can calculate the anomolous or magnetic declination for any place on earth and at any point in time. If you need to make adjustments, we suggest that you use this calculator to determine the declination for the area in question. Otherwise determine your heading using a reliable map.

8. Recording Extent

The extent is a measure of the size of the area within which collecting events or observations occurred for a given locality. Assuming the locality is recorded as a coordinate, the extent is the distance from that point to the furthest point where collecting or observations occurred in that locality. Extent has not traditionally been recorded with collecting activities, but can be

⁴ National Geophysical Data Center. 2004. [Estimated Value of Magnetic Declination](#).

important where activities have taken place over a small range, along a transect, or over an area (for example it is common to record bird observations over a 2 ha area).

Collecting events or observations often take place in an area described collectively by a single locality (e.g., within 1 km of the place described in the recorded locality). Without a measure of the potential deviation from the point provided, a user of the data usually has no way of knowing how specific the locality actually is. The extent is a simple way to alert the user that, for example, all of the specimens collected or observations made at the stated coordinates were actually within an area of up to 0.5 miles from that point. It can be quite helpful at times to include in your field notes a large-scale map of the local vicinity for each locality, marking the area in which the collecting and observations occurred.

Hint: A 1 km linear trap line for which the coordinates refer to the center has an extent of 0.5 km. A 2ha area where the coordinates are given at the center of a circle has an extent of ~80 m.

9. Recording Year of Collection

The year a collection was made can often affect the georeferencing of a location. Towns, roads, counties, and even countries can change names and boundaries over time. Rivers and coastlines can change position, billabongs and ox-bow lakes can come and go, localities (such as towns) can change size and shape, and areas of once pristine environment may become farmland or urban areas. Dated maps may no longer represent the current situation. The date is an important characteristic of the collection and must be taken into account when determining a georeference.

Example: “Collecting localities along the Alaska Highway are frequently given in terms of milepost markers; however, the Alaska Highway is approximately 40 km shorter than it was in 1942 and road improvements continue to re-route and shorten it every year. Accurate location of a milepost, therefore, would require cross-referencing to the collecting date. To further complicate matters, Alaska uses historical mileposts (calibrated to 1942 distance), the Yukon uses historical mileposts converted to kilometers, and British Columbia uses actual mileage (expressed in kilometers)”. (From Wheeler *et al.* 2001).

10. Documentation

Record the sources of all measurements. Minimally, include map name and scale, GPS model, the datum, the source for elevation data, the UTM Zone if using UTM coordinates, and the extent of the location or collecting event.

Using a GPS. For the best accuracy of a location determined by GPS it is important to document:

- The coordinates obtained from the GPS
- The datum
- The accuracy reported by the GPS
- Make of GPS receiver used

Note! Most GPS devices do not record accuracy with the waypoint data, but provide it in the interface showing current satellite conditions.

Note!: The accuracy reported by most GPS recorders is only a relative accuracy for the instrument on which it is read and not real accuracy. For many GPS recorders, the accuracy reported is almost always smaller than warranted.

Example:

Locality: “Modoc National Wildlife Refuge, 2.8 mi S and 1.2 mi E junction of Hwy. 299 and Hwy. 395 in Alturas, Modoc Co., Calif.”

Lat/Long/Datum: 41.45063, -120.50763 (WGS84)

Elevation: 1330 ft

GPS Accuracy: 24 ft

Extent: 150 ft

References: Garmin Etrex Summit GPS for coordinates and accuracy, barometric altimeter for elevation.

(From [MVZ Guide for Recording Localities in Field Notes](#))

11. Recording Data for Small Labels

An issue that often arises with insect collections is the problem of recording locality information on small labels. This should not be as big a problem as previously because new technologies allow for linking information on the label to a database (through bar codes, etc.) with the recording of basic information on the label. See Wheeler *et al.* (2001) on guidelines for preparing labels for terrestrial arthropods, but bear in mind the principles laid out in this document when preparing data for insect labels, especially the recording of datums, etc., which are not covered in that document.

12. New Technologies

A number of new technologies are beginning to make data recording in the field a lot easier. For example, a number of companies have recently released Personal Digital Assistants (PDAs) with built-in GPS receivers that can, depending on the type, record to a relatively high degree of accuracy. While these are excellent for recoding locality information in the field for later transfer to the database and for the preparation of labels, many do not include an exterior aerial for receipt of the satellite data and this is likely to reduce the accuracy of the recorded information. The lack of an exterior aerial makes the need for clear line of site for the satellites more important.

The use of Globally Unique Identifiers (GUIDs) for uniquely identifying individual objects and other classes of data (such as collections and observations) are under discussion. We recommend that these be followed once a stable system is implemented. Further information can be found on the TDWG⁵ and GBIF⁶ websites.

⁵ http://www.tdwg.org/TDWG_GUID.htm

⁶ <http://www.gbif.org>

Beginning the Georeferencing Process

1. Introduction

A number of issues must be addressed before one begins to georeference. It may appear to be a daunting task at the beginning, however there are many ways the process can be simplified and made more practical.

Managers and curators are sure to ask many of the following questions and more:

- How hard is this going to be?
- How long is it going to take?
- What proportion of my collection is already digitized?
- What is the current condition of the collection?
- What are the advantages and disadvantages of georeferencing the collection?
- How will the georeferenced data be used and by whom?
- What kind of expertise am I going to need?
- What supervision will be needed and who will do it?
- To what extent will I have to, or want to change my data model?
- How much is it going to cost and what resources are available for georeferencing?
- What tools exist to help me?
- Can I trust what comes out of these tools?
- How many data entry staff will I need?
- What training will I need to give my data entry staff?
- How much of the established best practices do I really need to follow?

This document will not answer all these questions, as many are institution specific, however, it should provide the answer to some, and provide the means of determining the others.

The first issue that will need to be addressed is the database management system:

- Will my current database cope or do I need to have it modified?
- How will I need to modify my user interface to make it easier for data entry operators to georeference?
- What is the most efficient way to go about data entry, including the georeferencing?

This document does not cover methods of general data entry. There are many ways that this may be conducted. These include direct entry from the label with the specimen or ledger brought to the computer; use of PDA's where the computer is brought to the specimen; the use of scanning or photographic (still or video) equipment to capture the label information so that the data entry operator can enter the information from a screen; or use of handwriting and OCR tools to capture the data, etc. Some of these methods are only just becoming practical, but you should make an active decision on the method that best suits your institution.

The next section will help you decide if your database will need modifying or not, and to what extent. It is often tempting to just include fields for the georeferenced coordinates and ignore any additional fields; however, you (or those who follow after you) are sure to regret taking such an option further down the line. The associated information on methods used to determine the georeference, and on the extent and uncertainty associated with the georeference, are very important pieces of information for the end user. Additionally, these are very important pieces of information for managing and improving the quality of your information.

Good examples of production systems that are well documented are the [Mountains and Plains Spatio-Temporal Database Informatics Initiative](#) (MaPSTeDI) program and the [Mammal Networked Information System](#) (MaNIS). It is worth looking at the processes these projects go through for georeferencing data.

2. The Resources Needed

Each institution will have needs for different resources in order to georeference their collections. The basics, however, include:

- A database and database software (we do not recommend the use of spreadsheets)
- Topographic maps (electronic, paper or both)
- Access to a good gazetteer – (many are available free via the Internet, either for downloading, or via on-line searching)
- Preferably internet access (as there are many resources on the Internet that will help in georeferencing and locating places)
- Suitable computer hardware

Further information on some of these requirements can be found on the MaPSTeDI site under [“What you Need”](#).

3. Fields to Include in your Database

One of the key aspects to efficient georeferencing is setting up a database correctly.

Some georeferencing projects (e.g., [MaPSTeDI](#)) use a separate working database for data entry operators so that the main data are not modified and day-to-day use of the database is not hindered. The data from the working database can be checked for quality, and then uploaded to the main database from time to time. Such a way of operating is institution dependant, and may be worth considering.

a. Determine what fields you need⁷

This step seems self-explanatory but it is surprising how often a database is created and finalized before it is determined exactly what the database is supposed to hold. The supervisors for the georeferencing process should be consulted before the database is created to ensure the required georeferencing fields are included in the data model from the outset. Be sure not to lump together dissimilar data into one field. Always atomize the data into separate fields where possible. For instance, if you are collecting latitude and longitude, your database should at least have a separate field for each. Finally, it is also appropriate to use this discussion to decide which fields the data entry operators should see when they are georeferencing. Fields such as date of collection, collector, specimen ID, and taxonomy are very helpful for georeferencing operators to see along with the more obvious locality data.

Note! When you are atomizing data on entry, always include a field or fields that record verbatim the original data so that atomization and other transformations can later be revealed and checked.

⁷ Modified from the MaPSTeDI Guidelines

<http://mapstedi.colorado.edu/GuideToGeoreferencing/Georeferencing1-3_SettingUpYourDatabase.html> .

b. Locality fields

What are the fields you need in your database to best store georeferencing information? This can perhaps best be divided into two parts, the first are those fields associated with the locality description. Many institutions are currently breaking down locality descriptions into their component parts, i.e., location name, distance and direction, etc., and include this information in separate fields in their databases. With the development of the [BioGeomancer](#) toolkit, however, and its automated parsing of natural language locality descriptions, this is now becoming redundant and unnecessary (see further discussion, below). If this break-up of locality information is done, it is important not to replace the free-text locality field (the data as written on the label or in the field notebook), but to add additional fields, as the written format of the description is often important, and this original information should never be over-written or deleted.

Other fields that may be important and useful to aid in georeferencing are:

- date last modified
- township/section/range/Local Government Area/county/state/country
- elevation
- date of collection
- remarks.

A reference worth checking before developing your own data base system is the *Herbarium Information Standards and Protocols for Interchange of Data* (Conn 1996, 2000), which although set up for herbaria, is applicable to most natural history collection data.

c. Georeferencing fields

The second set of fields are those fields actually associated with the georeference, and the georeferencing process. It is recommended, for best practice in georeferencing, that the following fields⁸ be added to your database as a minimum. These are in addition to other fields your database may already have, such as Latitude_Degrees, Latitude_Minutes, Latitude_Seconds, etc. Some databases include a user interface to the database that allows data to be entered as degrees, minutes, second, but then translates it to decimal degrees on entry into the database. If this is the case, then both sets of georeferences should be stored, with the decimal degrees used for data exchange. See also the [Geospatial Element Definitions Extension to Darwin Core](#) (TDWG 2005).

Field	Comments
Decimal Latitude	See <i>Glossary</i> for definition. Positive numbers are north of the equator and are less than or equal to 90, while negative values are South of the Equator and are greater or equal to -90. <u>Example</u> : -42.5100 degrees (which is roughly the same as 42° 30' 36" S).

⁸ From the Museum of Vertebrate Zoology Georeferencing Guidelines <<http://manisnet.org/GeorefGuide.html>>

Field	Comments
Decimal Longitude	See <i>Glossary</i> for definition. Positive values are East of the Greenwich Meridian and are less than or equal to 180, negative values are West of the Greenwich Meridian and greater than or equal to -180. <u>Example:</u> -122.4900 degrees (which is roughly the same as 122° 29' 24" W).
Geodetic Datum	The geometric description of a geodetic surface model (e.g., NAD27, NAD83, WGS84). Datums are often recorded on maps and in gazetteers, and can be specifically set for most GPS devices so the waypoints match the chosen datum. Use "not recorded" when the datum is not known. [See separate discussion on datums in this document].
Maximum Uncertainty Estimate	The upper limit of the distance from the given latitude and longitude describing a circle within which the whole of the described locality must lie.
Maximum Uncertainty Unit	The unit of length in which the maximum uncertainty is recorded (e.g., mi, km, m, and ft). Express maximum uncertainty distance in the same units as the distance measurements in the locality description.
Verbatim Coordinates	The original (verbatim) coordinates of the raw data before any transformations were carried out.
Verbatim Coordinate System	The coordinate system in which the raw data were recorded. If data are being entered into the database in Decimal Degrees, for example, the geographic coordinates of the map or gazetteer used should be entered (e.g., decimal degrees, degrees-minutes-seconds, degrees-decimal minutes, UTM coordinates).
Georeference Verification Status	A categorical description of the extent to which the georeference and uncertainty have been verified to represent the location and uncertainty for where the specimen or observation was collected. This element should be vocabulary-controlled. Examples: 'requires verification', 'verified by collector', 'verified by curator', 'not verified', etc.
Georeference Validation	Shows what validation procedures have been conducted on the georeferences – for example various outlier detection procedures, revisits to the location, etc. Relates to Verification Status.
Georeference Protocol	A reference to the method(s) used for determining the coordinates and uncertainty estimates (e.g., "MaNIS Georeferencing Calculator").
Georeference Sources	The reference source (e.g., the specific map, gazetteer, or software) used to determine the coordinates and uncertainties. Such information should provide enough detail so that anyone can locate the actual reference used (e.g., name, edition or version, year). Map scales should be recorded in the reference as well (e.g., USGS Gosford Quad map 1:24000, 1973).
Spatial Fit	A measure of how well the geometric representation matches the original spatial representation and is reported as the ratio of the area of the presented geometry to the area of the original spatial representation. A value of 1 is an exact match or 100% overlap. This is a new concept for use with biodiversity data, but one that we are recommending here. [See section on Spatial Fit later in this document].

Field	Comments
Georeference Determined By	The person or organization making the coordinate and uncertainty determination.
Georeference Determined Date	The date on which the determination was made.
Georeference Remarks	Comments on methods and assumptions used in determining coordinates or uncertainties when those methods or assumptions differ from, or expand upon, the methods referenced in the Georeference Protocol field.

d. Ecological data

The georeferencing portion of an ecological data collection should be treated in a similar way to specimen and observation data. Often ecological data are recorded using a grid, or transect, etc., and may have a starting locality and an ending locality as well as start time and end time. Sometimes the center of the transect is used as the locality, and half of the length of the transect used for the extent. The uncertainty is then calculated as for other data. If the data are recorded in a grid, then the locality is recorded as the center of the grid, and the extent from that position to the furthest extremity (i.e., the corner) of the grid. These data should be in addition to the recorded locality data, especially where many different fields are used to record the original data. See comments in Appendix.

e. Applying constraints

One of the key ways of making sure that data are as clean and accurate as possible is to assure that data cannot be put in the wrong field and that only data of a particular type can be put into each field. This is done by applying constraints on the data fields – for example, only allowing values between +90 and –90 in the decimal_latitude field. Many of the errors found when checking databases are needless errors – errors that should not be allowed to occur if the database had been set up correctly in the first instance.

With ecological or survey data etc., one could set boundary limits between the starting locality and ending locality. For example, if your methodology always uses 1 km or shorter transects, then the database could include a boundary limit that flagged whenever an attempt was made to place these two points more than 1 km apart.

4. User Interfaces

Good user-friendly interfaces are essential to make georeferencing efficient and fast, and to cut down on operator errors. The layout should be friendly, easy to use, and easy on the eyes. Where possible (and the software allows it) a number of different views of the data should be presented. These views can place emphasis on different aspects of the data and help the data entry operator's efficiency by allowing different ways of entering the data and by presenting a changing view for the operator, thus cutting down on boredom.

In the same way, macros and scripts can help with automated and semi-automated procedures, reducing the need for tedious (and time-consuming) repetition. For example, if data are being entered from a number of collections by one collector, taken at the same time from the same location, the information that is repeated from record to record should be able to be entered using just one or two key strokes.

5. Using Standards and Guidelines

Standards, standard methodologies, and guidelines can help lead to consistency throughout the database and cut down considerably on errors. A set of standards and guidelines should be established at the start of the process and before any georeferencing begins. They should remain flexible enough to cater for new data and changes in processes over time. Standards and guidelines in the following areas can improve the quality of the data and the efficiency of data entry. It is hoped that this document will provide guidelines for many of these. They include:

- Units of measure. Use a single unit of measure in interpreted fields. For example, do not allow a mixture of feet and meters in elevation and depth fields. Irrespective of this, the original units and measurements should be retained in a verbatim field.
- Methods and formats for determining and recording uncertainty and extent.
- Degree of accuracy in determining points where known. (For much legacy data, this will not be determinable).
- Fields that must be filled in (i.e. required fields).
- Format for recording coordinates (i.e., for lat/long, degrees/minutes/seconds, degrees/decimal minutes, or decimal degrees).
- Original source(s) of place names.
- Dealing with typos and other errors in the existing database.
- Number of decimal places to keep in decimal numbers.
- How to deal with “null” values as opposed to zero values (some databases have problems with this).
- How to deal with mandatory fields that cannot be filled in immediately (for example, because a reference has to be found). There may be need for something that can be put in the field that can allow the database to be filed and closed, but that flags that the information is still required.
- What data validation is to be carried out before a record can be considered complete?

Determining these standards and documenting them can help you to maintain them as well as assist you in training and data quality recording. They should form part of the institution’s own georeferencing best practice manuals.

6. Choosing a Methodology

Institutions and many experienced georeferencers develop their own preferences for the order in which they georeference. This may be determined by the nature of the data, the way specimens are stored or documented or on the general preference of the operator.

The MaPSTeDI project makes the following recommendations. Note that these will not suit every institution, but may provide a guide:

Georeferencing Procedures

Step 1 - Locate and plot the locality point

The actions involved in this step are described in Finding Coordinates.

Step 2 - Assign a confidence value to the locality

The actions involved in this step are described in Assigning Confidence Values.

Step 3 - Record the georeferenced locality data

This is an important but often under-appreciated step. Most of the mistakes in georeferenced data come from incorrectly recorded data. It is important that all required database fields be filled in as completely as possible in the correct format. The database administrator should place constraints upon some fields to force correct format.

Step 4 - Document the georeferencing rationale for each record

This step is critical because it documents the decision making process for each georeferenced record. For problem records, as well as confusing or detailed records, this information is very important to permit quality checking personnel and museum database users to understand the rationale behind the locality point and confidence value selection. This information also serves as a daily log which permits georeferencing personnel to communicate ideas and report problems. This documentation should be databased with the georeferenced data. If databasing this information is not possible due to database software limitations, it should be kept in electronic documents.

Step 5 - Mark record for further review, if necessary

If the locality cannot be found or is confusing, it should be marked for review by quality checking personnel. This can occur in the database itself or however it is most convenient, but the georeferencer should attempt to complete the record if possible to expedite the quality checking process. The georeferencer should also collect as much relevant locality data as possible to aid the quality checker.

From [MaPSTeDI](#) (2004).

a. Sorting records for batch georeferencing

Another set of questions revolves around whether you are best georeferencing each record as you enter the data into the database or if it is better to georeference in a batch after the information on the label has been entered. There are arguments for each method, and again the circumstances of your institution should dictate the best method for you. If your data are stored taxonomically and not geographically (as is the case in the majority of instances) it is often best to georeference in a batch mode by sorting the locality data electronically, and in this way you can deal with many records on one map sheet or area at a time and not be jumping back and forth between map sheets. In other cases, there may be less wear and tear on collections, you may wish to database collections as they are received and before distributing duplicates, or sending on loan, or there may be other good practical reasons to georeference as you go. One advantage of georeferencing as you go is that you may be able to do all the collections of one collector at a time, and virtually follow his/her path, thus reducing errors from not knowing which of several localities may be correct.

Often there is value in georeferencing in batch (tools such as [BioGeomancer](#), work better this way) or in collaboration (MaNIS and MaPSTeDI found that collaborative georeferencing resulted in great efficiency gains), but then reviewing the records using collector and date, or looking at the records taxonomically to check for outliers, and other such data quality flags, afterwards. It usually boils down to what is the best method for your institution, but first, you should consider each of the alternatives before deciding which to use.

The data, once entered into the database, may be sorted using the locality field itself, or some other field such as region, state, nearest named place, etc. You may be able to sort the data into:

- map squares (C-squares⁹ often used for marine data, map sheets, UTM zones, etc.)
- geographic regions (country, state, local government area, etc.)
- named place (town, river)

⁹ C-Squares <<http://www.marine.csiro.au/csquares/about-csquares.htm>>

- collector, collector number, and date of collection.

Note! Major efficiency gains can usually be made by georeferencing in batch mode. Consider also, georeferencing collaboratively with other researchers or institutions with similar goals and complementary resources.

b. Using previously georeferenced records

It may be possible to use a look-up system that searches the database for similar localities that may have already been georeferenced. For example, if you have a record with the locality “10 km NW of Campinas”, you can search the database for all records with locality “Campinas” and see if any records that mean the same thing as “10 km NW of Campinas” have been georeferenced previously.

An extension of this method could use the benefits of a distributed data system such as the [Global Biodiversity Information Facility](#) (GBIF) Portal. A search could be conducted to see if the locality had already been georeferenced by another institution. At present, we quite often find that duplicates of the one collection have been given different georeferences by different institutions. The problem is knowing which of the several georeferences may be the correct one, and one needs to put a lot of faith in another institution’s georeferencing methodologies and accuracy determination. This gives strength to the arguments for good documentation with georeferencing, collaboration, and the recording of maximum uncertainty.

Care! This method can add error, if a mistake was made the first time, it will be perpetuated through all later instances.

c. Using BioGeomancer

The BioGeomancer Consortium has developed an online workbench, web services, and desktop applications that will provide georeferencing for collectors, curators and users of natural history specimens, including software tools to allow natural language processing of archival data records that were collected in many different formats and languages. The BioGeomancer Workbench will be launched in September 2006 and is founded on the pioneering efforts of four existing applications, [BioGeoMancer Classic](#), [GEOLocate](#), [DIVA-GIS](#), and the [MaNIS Georeferencing Calculator](#), as well as a number of innovations such as machine learning, spatial data editing, data validation and outlier detection.

BioGeomancer allows the submission of locality descriptions, either singly or in batch mode, and reports back the georeference, along with information on uncertainty. It also passes the data (and other data submitted by the user) through a number of validation tests to check for possible errors in already georeferenced data and to provide further information where several options exist from the locality information.

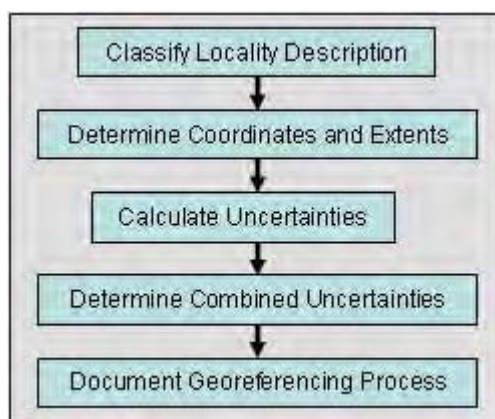
7. Data Entry Operators

The choice and training of data entry operators can make a big difference to the final quality of the georeferenced data. As mentioned earlier, the provision of good guidelines and standards can help in the training process and allow for data entry operators to reinforce their training over time. One of the greatest sources of georeferencing error is the data entry process. It is important that this process be made user-friendly, and be set up so that many errors cannot occur (e.g., through the use of pick lists, field constraints, etc.).

Georeferencing Legacy Data

By far the most difficult issue in georeferencing primary species occurrence data is the massive amount of legacy data held in the world's museums, herbaria, universities, etc. Most modern collectors are now using GPSs or large scale maps to locate their collection events, and thus most of the new data entering institutions already include georeferences. Most museums beginning to database their collections, however, are faced with the massive task of georeferencing the huge backlog of data in their collections, much of it with very little or vague location information. This document aims to assist these institutions with georeferencing their legacy data.

Wieczorek *et al.* (2004) identified five key steps to georeferencing. These have been modified slightly here to include:



Note! These steps should be considered in conjunction with the Appendix to this document.

Refer to the original document for a detailed explanation. We have extracted key points and elaborated on those below.

1. *Classifying the Locality Description*

Locality descriptions of primary species occurrence data encompass a wide range of content in a vast array of formats, but mostly are cited as a free text description. There are a limited number of categories that locality descriptions can be placed into for georeferencing purposes. The locality type determines the best method of calculating coordinates and uncertainties (see Appendix).

A locality description can contain multiple clauses and can match more than one category. If any one of the parts falls into one of the four categories, 'dubious' 'cannot be located', 'demonstrably inaccurate', or 'captive or cultivated' (see Appendix), then the locality should not be georeferenced. Instead, an annotation should be made to the locality record giving the reason why it is not being georeferenced.

If the locality description does not fall into one of those four categories, the most specific part of the locality description should be used for georeferencing. For example, a locality written as

'bridge over the St. Croix River, 4 km N of Somerset'

should be georeferenced based on the ***bridge*** rather than on ***Somerset*** as the named place with an offset at a heading. The locality should be annotated to reflect that the bridge was the

locality that was georeferenced. If the more specific part of the locality cannot be unambiguously identified, then the less specific part of the locality should be georeferenced and annotated accordingly.

2. Finding the Latitude and Longitude

As discussed elsewhere in this document, geographic coordinates can be expressed in a number of different coordinate systems (decimal degrees, degrees minutes seconds, degrees decimal minutes, UTM, etc.). Conversions can be made readily between coordinate systems, but decimal degrees provide the most convenient coordinates to use for georeferencing for no more profound a reason than a locality can be described with only two attributes - decimal latitude and decimal longitude (Wieczorek 2001). Decimal Degrees are also the coordinate system used in most Geographic Information Systems (GIS).

The first step in determining the coordinates for a locality description is to identify the most specific named place within the description. Coordinates may be retrieved from gazetteers, geographic name databases, maps, or from other locality descriptions that have coordinates. We use the term '*feature*' to refer to not only traditional features, but also to places that may not have proper names, such as road junctions, stream confluences, highway mile pegs, and cells in grid systems (e.g., townships). The source and precision of the coordinates should be recorded so that the validity of the georeferenced locality can be checked. The original coordinate system and the geodetic datum should also be recorded. This information helps to determine sources and degree of maximum uncertainty, especially with respect to the original coordinate precision.

3. Using Offsets

An offset is a displacement from a reference point, named place, or other feature, and is generally accompanied by a direction (or heading). Some locality descriptions give a method for determining the offset ('by road', 'by river', 'by air', 'up the valley', etc.). In such cases, follow the path designated in the description using a map with the largest available scale to find the coordinates of the offset from the named place. It is sometimes possible to infer the offset path from additional supporting evidence in the locality description. For example, the locality

'58 km NW of Haines Junction, Klwane Lake'

suggests a measurement by road since the final coordinates by that path are nearer to the lake than going 58 km NW in a straight line. At other times, you may have to consult detailed supplementary sources, such as field notes, collectors' itineraries, diaries, or sequential collections made on the same day, to determine this information.

4. Finding the Extent

Every named place occupies a finite space, or 'extent'. The extent is usually measured as the distance from the geographic center of the shape that defines the feature, to the furthest extremity of that shape.

If the locality described is an irregular shape (e.g., a winding road or river), there are two ways of calculating the coordinates and determining the extent. The first is to measure along the vector (line) and determine the mid point as the location of the 'named place'. This is not always easy, so the second method is to determine the geographic center (i.e., the midpoint of the extremes of latitude and longitude) of the named place. This method describes a point

where the uncertainty due to the extent of the named place is minimized. The extent is then determined as the distance from the determined position to the furthest point at the extremes of the vector. If the geographic center of the shape is used and it does not lie within the locality described (e.g., the geographic center of a segment of a river does not actually lie on the river), then the point nearest the geographic center that lies within the shape is the preferred reference for the named place and represents the point from which the extent should be calculated.

Many localities are based on named places that have changed in size over time; current maps might not reflect the extents of those places when specimens were collected. If possible, extents should be determined using maps contemporary with the events. In most cases, the current extent of a named place will be greater than its historical extent.

5. Calculating Uncertainties

Calculating uncertainties in georeferenced data provides a key provision in determining the data's fitness for use and thus their quality. There are many methods of determining maximum uncertainty; however most of these are complicated, difficult to simply record in most current natural history databases, and are often more sophisticated than necessary for the level of data being used. Over time, it is likely that the recording of uncertainty will be by way of geographic polygons; however, at this stage we recommend the use of a simple point-radius method (see Wieczorek *et al.* 2004) to record the error. The point-radius method is designed to not underestimate the true error. The introduction of polygons will allow, for example, clipping a circle where it overlaps the ocean for terrestrial data, and thereby provide a much more accurate representation of the locality.

Whenever subjectivity is involved, it is preferable to overestimate the maximum error or uncertainty. The following six sources of uncertainty are the most common encountered and these are elaborated below and in the Appendix:

- the extent of the locality
- unknown datum
- imprecision in distance measurements
- imprecision in direction measurements
- imprecision in coordinate measurements
- map scale.

a. Calculating uncertainties due to an unknown datum

Seldom do natural history collections have geographic coordinates recorded together with geodetic datum information. Even with modern collections using a GPS to record coordinates, the geodetic datum is typically ignored. A missing datum reference, however, introduces ambiguity, which varies geographically and adds greatly to the error inherent in the georeferencing.

It is important to record the datum used for the coordinate source (GPS, map sheet, gazetteer) if it is known, or to record the fact that it is not known.

Differences between datums may cause an error in true location from a few centimeters to around 1000 meters (US Navy *n. dat.*), or even, in some extreme instances, up to 3.552 km (Wieczorek *et al.* 2004). Some known average and/or maximum differences between datums

are cited in Table 1. Note that the difference between datums is not a linear relationship and they do not always vary in the same direction. For example, the difference between NAD27 and WGS84 in the conterminous USA varies between 0 and 104 m (Wieczorek *et al.* 2004).

Datum from	Region or Location	Datum to	Difference
AGD66	Australia	AGD84	Max \pm 0-5 m
AGD66/84	Australia	GDA94	Max \pm 200 m
AGD66/84	Australia	WGS84	Max \pm 200 m
GDA94	Australia	WGS84	Max \pm <1 m
NAD 1983	North America	WGS84	Max \pm <1 m
NAD27	North America	WGS84	Max \pm 200 m
NAD 27	Contiguous USA	WGS84	Max \pm 105 m
NAD 27	Aleutian Islands, Alaska	WGS84	Max \pm 235 m
NAD 27	Hawaii	WGS 84	~ 500 m
TOKYO	Japan	WGS84	Max \pm 750 m
ED-50	Europe	WGS84	Max \pm 175 m
ARC-50	Africa	WGS84	Max \pm 265 m
INDIAN 1975	Bangkok, Thailand	WGS84	~ 405 m
INDIAN 1956	Delhi, India	WGS84	~ 135 m
INDIAN 1956	Mumbai, India	WGS84	~ 120 m
HONG KONG 1973	Hong Kong	WGS84	~ 320 m
LUZON	Manila, The Philippines	WGS84	~ 225 m
TOKYO-KOREA	Seoul, South Korea	WGS84	~380 m
KERTAU 1948	Singapore	WGS84	~190 m

Table 1: Shows the maximum differences over total range, or approximate differences at a location for a number of common datums. Data derived from US Navy (*n. dat.*), Srivastava and Ramalingam (2006) and Wieczorek *et al.* (2004). All except the very small values have been rounded to the nearest 5 m.

b. Calculating uncertainty from distance

Precision can be difficult to gauge from a locality description as it is seldom, if ever, explicitly recorded. Further, a database record may not reflect, or may reflect incorrectly, the precision inherent in the original measurements, especially if the locality description in the database has undergone normalization, reformatting, or secondary interpretation of the original description.

There are a number of ways of calculating uncertainty from distances. In this document, we have taken a conservative approach. The form in which a distance is written can often give an indication of the precision and hence the uncertainty. One method is to use half of the precision (for example, 10.5 mi N of Bakersfield could reasonably be expected to mean 10½ mi and thus be between 10.25 and 10.75 mi N, or 10.5 \pm 0.25 mi N of Bakersfield). The uncertainty in the measurement is thus 0.25 mi.

A second method, and that recommended here, is one proposed by Wieczorek *et al.* (2004) and assumes that many records have undergone a certain amount of interpretation or transformation when being entered into the database, and thus a record of 10¼ mi may be entered into the database as 10.25 mi. The precision implied in the value 10.25 is thus a false precision (see glossary) and should not be assumed to be between 10.24 and 10.26. The method of Wieczorek *et al.* (2004) bases the estimate of uncertainty on the fractional part of the distance – i.e. calculated by dividing 1 by the fractional denominator. Thus:

- for 9 km, the fraction is 1/1 and thus the uncertainty estimate is 1 km;
- for 9.5 km, the fraction is ½ and the uncertainty estimate 0.5km;
- for 9.25 km, the fraction is ¼ and the uncertainty estimate 0.25 km;
- for 9.6 km, the fraction is 1/10 and the uncertainty estimate 0.1 km.

For distance measurements which are positive integer powers of 10, the uncertainty estimate is based on 0.5 times ten to that power (see Table 2).

A third method, suggested by Frazier *et al.* (2004), is for distances that are given as multiples of 10, or fractions of 100 such as 25 and 75. This method recommends using 15% of the distance as the uncertainty. Thus, for 10 km, the uncertainty would be 1.5 km; and for 75 km it would be 11.25 km. This gives a smaller uncertainty than recommended by Wieczorek *et al.* for distances between 10 and 30 km, and a greater value for distances between 40 and 90 km (Table 2).

Example	Uncertainty (Wieczorek <i>et al.</i> 2004)	Uncertainty (Frazier <i>et al.</i> 2004)
10.6 km N of Bakersfield	0.1 km	
10.5 mi N of Bakersfield	0.5 mi	
10 km N of Bakersfield	5 km	1.5 km
30 km N or Bakersfield	5 km	4.5 km
140 mi N of Bakersfield	5 mi	21 mi
200 mi N of Bakersfield	50 mi	30 mi
2000 m N of Bakersfield	500 m	300 m

Table 2. Calculating uncertainty using the precision in a distance recording

Precision can also be masked or lost when measurements are converted, such as from feet to meters, or from miles to kilometers.

Care! Be careful that the value you are using for precision when calculating the uncertainty is a true precision and not a false precision. For example, converting a collector’s recording of 16 miles (with a precision of 1 mile) to 25.6 km (with a precision of 0.1 km) leads to a level of precision that is more than 10 times as precise as the original.

Note! Further details of calculations used to determine uncertainties from distance precision can be found in [Wieczorek \(2001\)](#) and Wieczorek *et al.* (2004)

c. Calculating uncertainties from extents of localities

The extents of named places are an important source of uncertainty. Points of reference for named places may change over time – post offices and courthouses are relocated, towns change in size, the courses of rivers change, etc. Moreover, there is no guarantee that the collector paid attention to any particular convention when reporting a locality as an offset from a named place. For example,

‘4 km E of Bariloche’

may have been measured from the post office, the civic plaza, or from the bus station on the eastern edge of town, or anywhere else in Bariloche. When calculating an offset, we generally have no way of knowing where the collector started to measure the distance.

We recommend uncertainty be determined by measuring the distance from the point marked by the coordinates to the point in the named place furthest from those coordinates. The magnitude of the uncertainty will be smallest if the coordinates mark the geographic center of the named place and the maximum uncertainty is then the distance from that point to the furthest point in the locality. In most cases, the current extent of a named place will be greater than its historical extent and the uncertainty may be somewhat overestimated if current maps are used. When documenting the georeferencing process, it is recommended that the named place, its extent, and the source of the information all be recorded.

d. Calculating uncertainty from direction

The calculation of uncertainty from the precision in which a direction is recorded depends on distance from the reference point. The uncertainty will increase as one moves further from the source. For simple calculations of precision due to direction – see Table 3.

Note! The uncertainty due to directional imprecision increases with distance, so it can only be calculated from the combination of distance and direction (see below).

Precision	Interpretation	Example	Directional Uncertainty
N	Between NW and NE	10.6 km N of Bakersfield	45°
NE	Between NNE and ENE	10.5 mi NE of Bakersfield	22.5°
NNE	Between N of NNE and E of NNE	10 km NNE of Bakersfield	11.25°

Table 3. Calculating uncertainty using the precision of the recorded direction (derived from Wieczorek *et al.* 2004).

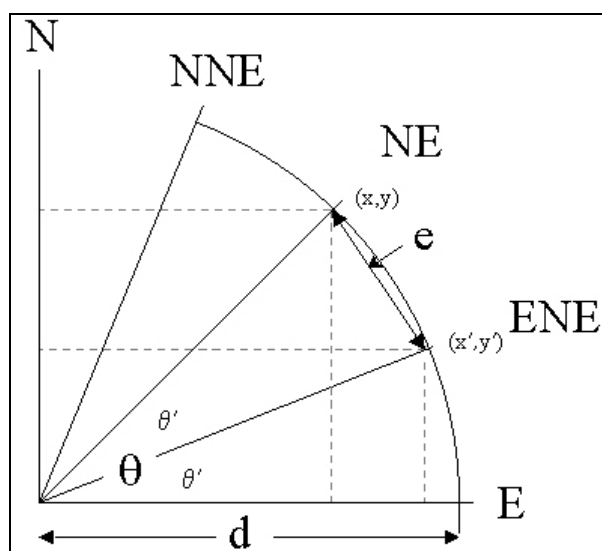


Fig. 1. A simple diagram showing directional precision where $x = d \cos(\theta)$, $y = d \sin(\theta)$, $x' = d \cos(\theta')$, and $y' = d \sin(\theta')$. From Wiczorek *et al.* (2004).

Using the example

'10 km NE of Bakersfield'

if we ignore distance imprecision, the uncertainty due to the direction imprecision (Figure 1) is encompassed by an arc centered 10 km (**d**) from the center of Bakersfield (at **x,y**) at a heading of 45 degrees (**θ**), extending 22.5 degrees in either direction from that point. At this scale the distance (**e**) from the center of the arc to the furthest extent of the arc (at **x',y'**) at a heading of 22.5 degrees (**θ'**) from the center of Bakersfield can be approximated by the Pythagorean Theorem.

$e = \sqrt{(x' - x)^2 + (y' - y)^2}$ the uncertainty in the above example is 3.90 km

This shows just one simple example. For details and formulae for calculating more complicated uncertainties, see [Wiczorek \(2001\)](#) and Wiczorek *et al.* 2004. Because of the complicated nature of these calculations, it is often best to use the [MaNIS Georeferencing Calculator](#) - see discussion below.

e. Calculating uncertainty from coordinate precision

Geographic coordinates should always be recorded using as many digits as possible; the precision of the coordinates should be captured separately from the coordinates themselves, preferably as a distance, which conserves its meaning regardless of location and coordinate transformations. Recording coordinates with insufficient precision can result in unnecessary uncertainties. The magnitude of the uncertainty is a function of not only the precision with which the data are recorded, but also of the datum and the coordinates themselves. This is a direct result of the fact that a degree does not correspond to the same distance everywhere on the surface of the earth.

Table 4 shows examples of the contributions to uncertainty for different levels of precision in original coordinates using the WGS84 reference ellipsoid. Calculations are based on the same degree of imprecision in both coordinates and are given for several different latitudes. Approximate calculations can be made based on this table, however, more accurate

calculations can be obtained using the [MaNIS Georeferencing Calculator](#) (see further discussion, below).

From Table 4, it can be seen that an observation recorded in degrees, minutes, and seconds (DMS) has a minimum uncertainty of between 32 and 44 meters.

Precision	0 degrees Latitude	30 degrees Latitude	60 degrees Latitude	85 degrees Latitude
1.0 degree	156,904 m	146,962 m	124,605 m	112,109 m
0.1 degree	15,691 m	14,697 m	12,461 m	11,211 m
0.01 degree	1,570 m	1,470 m	1,246 m	1,121 m
0.001 degree	157 m	147 m	125 m	112 m
0.0001 degree	16 m	15 m	13 m	12 m
0.00001 degree	2 m	2 m	2 m	2 m
1.0 minute	2,615 m	2,450 m	2,077 m	1,869 m
0.1 minute	262 m	245 m	208 m	187 m
0.01 minute	27 m	25 m	21 m	19 m
0.001 minute	3 m	3 m	3 m	2 m
1.0 second	44 m	41 m	35 m	32 m
0.1 second	5 m	5 m	4 m	4 m
0.01 second	1 m	1 m	1 m	1 m

Table 4. Table showing metric uncertainty due to precision of coordinates based on the WGS84 datum at varying latitudes. Uncertainty values have been round up in all cases. From Wieczorek (2001).

Care! *False precision* can arise when transformations from degrees minutes seconds to decimal degrees are stored in a database (see **Glossary** for expanded discussion).

Never use precision in a database as a surrogate for the coordinate uncertainty; instead, record the uncertainty explicitly, preferably as a distance.

Note! Details of calculations used to determine uncertainties in coordinate precisions can be found in [Wieczorek \(2001\)](#) and Wieczorek *et al.* (2004).

Example:

Lat: 10.27° **Long:** -123.6° **Datum:** WGS84

In this example, the lat/long precision is 0.01 degrees. Thus, latitude error = 1.1061 km, longitude error = 1.0955 km, and the uncertainty resulting from the combination of the two is 1.5568 km.

f. Calculating uncertainty by reading off a map

One of the most common methods of finding coordinates for a location is to estimate the location from a paper map. Using paper maps can be problematic and subject to varying degrees of inaccuracy. Unfortunately, the accuracy of many maps, particularly old ones, is undocumented. Accuracy standards generally explain the physical error tolerance on a printed map, so that the net uncertainty is dependent on the map scale. Map reading requires a certain level of skill in order to determine coordinates accurately, and different types of maps require different skills. Challenges arise due to the coordinate system of the map (latitude and longitude, UTM, etc.), the scale of the paper map, the line widths used to draw the features on the maps, the frequency of grid lines, etc.

The accuracy of a map depends on the accuracy of the original data used to compile the map, how accurately these source data have been transferred onto the map, and the resolution at which the map is printed or displayed. For example, USGS maps of 1:24,000 and 1:100,000 are different products. The accuracy is explicitly dependent on scale but is due to the different methods of preparation. When using a map, the user must take into account the limitations encountered by the map maker such as acuity of vision, lithographic processes, plotting methodologies, and symbolization of features (e.g., line widths) (NOAA 2001).

With paper topographic maps, drawing constraints may restrict the accuracy with which lines are placed on the map. A 0.5 mm wide line depicting a road on a 1:250,000 map represents 125 meters on the ground. To depict a railway running beside the road, a separation of 1-2 mm (250-500 meters) is needed, and then the line for the railway (another 0.5 mm or 125 meters) makes a total of 500-750 m as a minimum representation. If one uses such features to determine an occurrence locality, for example, then minimum uncertainty would be in the order of 1 km. If thicker lines were used, then appropriate adjustments would need to be made (Chapman *et al.* 2005).

Note! A digital map is never more accurate than the original from which it was derived, nor is it more accurate when you zoom in on it. The accuracy is strictly a function of the scale and digitizing errors of the original map.

Table 5 shows the inherent accuracy of a number of maps at different scales. The table gives uncertainties for a line 0.5 mm wide at a number of different map scales. A value of 1 mm of error can be used on maps for which the standards are not published. This corresponds to about three times the detectable graphical error and should serve well as an uncertainty estimate for most maps.

Scale of Map	Map Horizontal Uncertainty (Geosciences Australia ¹⁰)	Map Horizontal Uncertainty (USGS ¹¹)	NIMA Product	NIMA Product Accuracy (US Navy) ¹²
1:1000	0.5 m	2.8 ft		
1:10,000	5 m	28 ft		
1:25,000	12.5 m	70 ft	City Graphic	>50 m
1:50,000	25 m	139 ft	Topo	50 m
1:75,000			Nautical	75 m
1:100,000	50 m	278 ft		
1:250,000	160-300 m	695 ft	JOG	250 m
1:500,000			TPC	1,000 m
1:1 million	500 m	2,777 ft	ONC	2,000 m

Table 5. Horizontal uncertainty and accuracy associated with a 0.5 mm line on maps of different scales.

The table uses data from several sources. The TOPO250K Map series is the finest resolution mapping that covers the whole of the Australian continent. It is based on 1:250,000 topographic data, for which Geoscience Australia (2003) defines the accuracy as “*not more than 10% of well-defined points being in error by more than 160 meters; and in the worst case, a well defined point is out of position by 300 meters*”. The USGS Map Horizontal Uncertainty is calculated from US Bureau of Budget (1947) which states that “*for maps on publication scales larger than 1:20,000, not more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; for maps on publication scales of 1:20,000 or smaller, 1/50 inch.*” These values need to be taken into account when determining the uncertainty of your georeference. The third set of values was obtained from the US Navy with reference to various NIMA¹³ (US National Image and Mapping Agency) products.

If you are using phenomena that do not have distinct boundaries in nature to determine a locality (such as soils, vegetation, geology, timberlines, etc.) then err vastly on the side of conservatism when determining an uncertainty value as such boundaries are seldom accurate, often determined at a scale of 1:1 million or worse and would have a minimum uncertainty of between 1 and 5 km. Also be aware that coastlines vary greatly at different scales (see Chapman *et al.* 2005) and rivers are often straightened on smaller scale maps, and can thus include uncertainties far greater than are generally recorded on maps whose accuracies are determined from “well-defined” points such as buildings, road intersections, etc. In addition,

¹⁰ Based on 0.5mm of accuracy per unit of scale, except for the 1:250,000 map series where the figure supplied with the data has been used.

¹¹ Derived from United States National Map Accuracy Standards (US Bureau of Budget 1947) <http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/nmas/NMAS647.PDF>

¹² Navigator of the Navy https://www.navigator.navy.mil/navigator/accuracy_0009.ppt

¹³ US National Image and Mapping Agency (NIMA) <http://erg.usgs.gov/nimamaps/>

coastlines and river paths can change greatly over time (Bannerman 1999) and thus the date of the map needs to be taken into account when determining uncertainty.

For elevation where contours are drawn on a map, the vertical uncertainty is usually described as being half of the contour interval.

Care! Care must be used when using a digital map that records the scale in the form of text (1:100,000, etc.) rather than by using a scale bar, as the resolution of the computer screen, and the level of zooming will change the apparent scale of the map being viewed. (It does not change the scale at which the map was prepared). This also applies to maps printed from a digital map. When preparing digital maps, always include scale as a scale bar and do not just record scale in textual form (e.g., 1:20,000).

g. Calculating combined uncertainties

When combining uncertainties from different sources, it is not as simple as taking the average or adding them together. Uncertainties inherent in the location of the named place, in its extent, in the direction of the offset, and the distance of the offset, are just four sources that need to be combined to get an overall uncertainty. A detailed discussion of the calculations involved can be found in [Wieczorek \(2001\)](#) and [Wieczorek *et al.* \(2004\)](#), and for a practical way of calculating uncertainties in locations, we recommend use of the [MaNIS Georeferencing Calculator](#). In the Appendix to this document, we provide a number of examples.

h. Using the MaNIS Georeferencing Calculator

The [MaNIS Georeferencing Calculator](#)¹⁴ (Figure 2), is a java applet created as a tool to aid in the georeferencing of descriptive localities such as those found in museum-based natural history collections. It was specifically designed for the Mammal Networked Information System ([MaNIS](#)) Project and has been adopted as well by both [HerpNet](#), [ORNIS](#), and other collaborative database initiatives.

The application makes calculations using the methods described in the [Georeferencing Guidelines](#) (Wieczorek 2001). We recommend its use generally by all natural history institutions to calculate uncertainty in location data without the need for a detailed understanding of the complicated underlying algorithms. The more institutions that use this one method, the more consistent will be the quality of data across and between institutions, making it easier for users to evaluate the quality of the data. We recommend reading both [Wieczorek \(2001\)](#) and the [MaNIS Georeferencing Calculator Manual](#) (Wieczorek 2002) for an understanding of the calculations involved and an understanding of how the calculator works.

The algorithms developed for the Georeferencing Calculator have also been incorporated in the the uncertainty calculations used in the BioGeomancer georeferencing tools. This too will serve to standardize the determination of this important attribute of data quality documentation.

¹⁴ MaNIS Georeferencing Calculator <<http://www.manisnet.org/gc.html>>

Version 020411

Georeferencing Calculator

Calculation Type:

Locality Type:

Step 3) Enter all of the parameters for the locality.

Coordinate Source:

Coordinate System:

Latitude: ° ' "

Longitude: ° ' "

Datum:

Coordinate Precision:

Offset Distance:

Extent of Named Place:

Distance Units:

Distance Precision:

Direction:

Decimal Latitude	Decimal Longitude	Maximum Error Distance	
35.37333	-118.84068	9.930	mi
degrees minutes seconds nearest second 1 mi 35.37333 -118.84068 (NAD27) North American			

Georef Calculator

Fig. 2. A snap shot of the MaNIS Georeferencing Calculator showing maximum uncertainty calculation for the locality: '10 mi E (by air) Bakersfield'. From Wiecezorek (2002).

6. Determining Spatial Fit

Spatial fit is a new georeferencing concept designed to allow for a measure of how well a given geometric representation matches the original spatial representation. This is useful when spatial transformations change the way a locality is represented, either to mask its detail, or to match an agreed upon schema for data sharing (such as fitting locations to a grid cell).

A spatial fit with a value of 1 is an exact match or 100% overlap. If the geometry given does not completely encompass the original spatial representation, then the spatial fit is zero (i.e., some of the original is outside the transformed version, which we interpret as not being a fit). If the transformed shape does completely encompass the original spatial representation, then the value of the spatial fit is the ratio of the area of the transformed geometry to the area of the original spatial representation. Special case: If the original spatial representation is a point and the geometry presented in not a point, then the spatial fit is undefined. The range of values of spatial fit is 0, 1, greater than 1, or undefined.

An example of the applicability of the spatial fit is where a point representing a terrestrial collection lies close to the coast, and the calculated uncertainty radius encompasses some marine area. In this case the Spatial Fit would be greater than 1 as it represents an area greater than the real uncertainty.

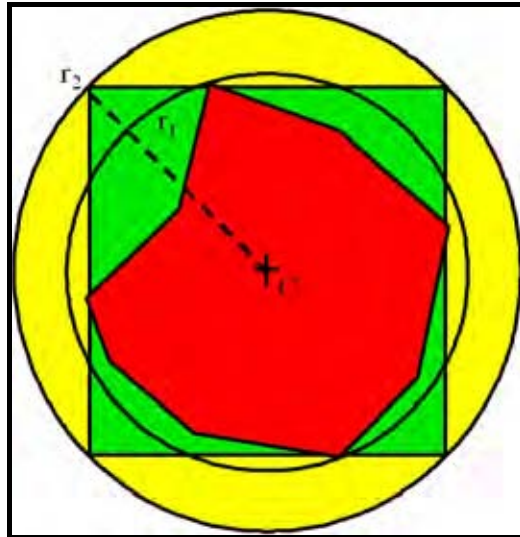


Fig. 3. A diagram illustrating the spatial fit of a number of locations that can be described by a polygon, a grid, or a point.

Figure 3 illustrates a few examples of the definition of spatial fit and these are elaborated below:

- 1) Suppose the original spatial representation of a locality was given by the **red polygon** with area **A**.

The spatial fit of the yellow circle would be	$(\text{PI} \cdot r_2^2) / A$
The spatial fit of the green bounding box would be	$(2 \cdot r_2^2) / A$
The spatial fit of the black circle (r_1) would be	$(\text{PI} \cdot r_1^2) / A$
The spatial fit of the red polygon would be	1
The spatial fit of the point C would be	0

- 2) Suppose the original spatial representation of a locality was given as the **green bounding box** with area $2 \cdot r_2^2$.

The spatial fit of the yellow circle would be	$(\text{PI} \cdot r_2^2) / (2 \cdot r_2^2)$
The spatial fit of the green bounding box would be	1
The spatial fit of the black circle (r_1) would be	0
The spatial fit of the red polygon would be	0
The spatial fit of the point C would be	0

- 3) Suppose the original spatial representation of a locality was given as the **black circle** with area $\text{PI} \cdot r_1^2$.

The spatial fit of the yellow circle would be	r_2^2 / r_1^2
The spatial fit of the green bounding box would be	0
The spatial fit of the black circle (r_1) would be	1
The spatial fit of the red polygon would be	0
The spatial fit of the point C would be	0

- 4) Suppose the original spatial representation of a locality was given as the **point C**.

The spatial fit of the yellow circle would be	Undefined
The spatial fit of the green bounding box would be	Undefined
The spatial fit of the black circle (r_1) would be	Undefined
The spatial fit of the red polygon would be	Undefined
The spatial fit of the point C would be	1

Maintaining Data Quality

Data that have been incorporated into the database and georeferenced need to be maintained and checked for quality. The quality checking process involves a number of steps, including receiving feedback from users, providing feedback to collectors, and running various validation tests. For more information on *data quality* and what it means for primary species collection data see Chapman (2005b). Two major principles associated with data quality and data cleaning are:

- Error prevention is preferable to error correction.
- The earlier in the information chain that you can detect an error, the cheaper it will be to correct it.

1. Feedback to Collectors

Maintaining the quality of the data may require giving feedback to others. For example, if you find that a particular collector is not recording his collection information correctly (e.g., not recording the datum with the georeference information), then you need to provide feedback to him so that future records have a lower level of error and thus a higher quality. See the earlier chapter on *Collecting and Recording Data in the Field*. Key issues that may require feedback to collectors include:

- Making sure the datum is recorded with all GPS readings
- Encouraging consistent use of a standard coordinate system (e.g., encourage collectors to use decimal degrees wherever possible)
- Recording localities in a consistent and clear manner
 - Using nearest named place and offsets
 - Recording ‘by road’ or ‘by air’
- Using a barometric altimeter for recording elevation.

2. Accepting Feedback from Users

Feedback from users can be one of the most valuable resources for maintaining the quality of one’s collections. For this to work, however, the institution needs to set up a good feedback mechanism. There needs to be a process whereby all feedback related to quality are checked and the results documented (see Chapman 2005a, b). Feedback may be from other institutions holding duplicates of some of your collections, from users who are carrying out analyses on large amounts of data and find records that are either wrongly georeferenced, or wrongly identified, or from users who are carrying out data quality checking on related records. All feedback is important, and should not be ignored. Checks carried out should also always be documented so that the same ‘error’ is not checked over and over again.

3. Data Checking and Cleaning

An important but often overlooked aspect to any georeferencing project is the checking of the georeferenced data that goes into the database. This aspect is often ignored because of lack of funds or personnel. However, because the point of any georeferencing project is to produce geographic coordinates linking a specimen to a place on a map or environmental data, it is important that the coordinates chosen are truly the best ones for the location. Not only does it improve the quality of data, but it also identifies trends and habits in georeferencing that may

need to be corrected. Often a graduate assistant, intern, or someone with more experience will do most of the quality checking.

a. Data entry

One of the major sources of error in georeferencing is at the stage of data entry. Errors can be reduced by the establishment of good data entry procedures – use of pick lists, field constraints, etc., to reduce the possibility of error. However, once these are in place and working, then regular checks need to be carried out on the data entry operators and on the process of data entry. Quality checking can take several forms, but we recommend that it use the following two taken from the [MapSTeDI Georeferencing Guidelines](#).

The first is to check the accuracy of the georeferencing. This process involves checking a certain number of each georeferencer's records. Based on various trials, it is recommended that the first 200 records that a new georeferencer completes be checked for accuracy. Not only is this initial checking beneficial to the accuracy of the data, but also it is essential to allow the georeferencer to improve and learn from making mistakes. If significant problems still exist after the initial 200 records, an additional batch of 100 records should be checked. After the quality checker, usually a highly experienced georeferencer, is satisfied with the new georeferencer's abilities, the quality checking is reduced to 10 randomly selected records out of every 100 completed. If more than two records are found to be incorrect within that 10, an additional 20 records should be checked. The quality checker may ask the georeferencer to redo the entire 100 if enough problems exist. After a period of few mistakes, the checking is reduced to five records for every 100 or at the quality checker's discretion.

To summarize:

- Initial 200 records should be checked. If problems remain, check groups of 100 until satisfied with georeferencer's abilities.
- Regular checks of 10 randomly selected records for every 100
- If more than 2 incorrect records, quality checker should check 20 more records and can ask georeferencer to redo entire 100.
- After awhile, the regular checks can be reduced to 5 records for every 100.

The second purpose of quality checking is to allow georeferencers to refer difficult or confusing records to the quality checker for help or advice. The quality checker will then resolve these 'problem records' as well as possible. Checking problem records can be like detective work. Historical records often have locality descriptions with named places that do not appear on modern maps or gazetteers. To find these localities, it is often necessary to consult several different sources of information. These sources include, but are not limited to catalog books, field notes, other records with similar localities, other collections, scientific and other publications, websites, online databases, specialty gazetteers, and historical maps. Bits of information from several places can often be used to establish the correct coordinates for a historical locality. In addition, some problem records do not make sense because of contradictions or missing or garbled information (see locality type categories in Appendix). These problem records may be the result of mistakes in data entry made in either the paper catalog or the database. It may also be necessary to consult the curatorial staff or even the original collector.

b. Data validation

Data validation (checking for errors) can be a time-consuming process, however, it is one of the most important processes you can carry out with your data. It is not practical to check every record individually, so the use of batch processing techniques and outlier detection procedures, etc. are essential. Fortunately, a number of these have been developed and are available in software products or on-line. Most of these are elaborated in the document [Principles and Methods of Data Cleaning. Primary Species and Species Occurrence Data](#) published by GBIF (Chapman 2005b) and the information therein will not be repeated here. We recommend that you download and use that document as an adjunct to this one.

There are many methods of checking for errors in georeferenced data. These can involve

- using external databases (collector's itineraries, gazetteers, etc.),
- checking against other fields in your own database (making sure the georeference falls within the correct state, country, region, etc.),
- using a GIS to look for records that fall outside polygon boundaries such as bioregions, local government areas,
- using statistical methods such as box plots, reverse jackknifing, cumulative frequency curves, and cluster analysis to identify outliers in latitude or longitude,
- using modelling software in conjunction with statistical analysis to identify outliers in environmental (e.g., climate) space.

Some of these techniques will shortly be available on-line through the [GBIF Portal](#), and the [BioGeomancer](#) website, and yet others are available through the stand-alone GIS software [DIVA-GIS](#) (Hijmans *et al.* 2005).

c. Making corrections

When making corrections to your database, we strongly recommend that you always add and never replace or delete. For this to happen you will usually require additional fields in the database. For example, you may have 'original' or 'verbatim' georeference fields in addition to the main georeference fields. Additionally, the database may require a number of 'Remarks' fields. Fields that can be valuable are those that describe validation checking that has been carried out – even (and often especially) if that checking has led to confirmation of the georeference. These fields may include information on what checks were carried out, by whom, when and with what results.

d. Truth in labelling

'Truth in Labelling' is an important consideration with respect to documenting data quality. This is especially so where data are being made available to a wider audience, for example, through the GBIF data nodes. We recommend that documentation of the data and their quality be up-front and honest. Error is an inescapable character of any dataset, and it should be recognized as a fundamental attribute of those data. All databases have errors, and it is in no-one's interest to hide those errors. On the contrary, revealing data actually exposes them to editing, validation and correction through user feedback, while hiding information almost guarantees that it remain dirty and of little long-term value.

4. Responsibilities of the Manager

It is important that the manager maintain good sets of documentation (guidelines, best practice documents, etc.), ensure that there are good feedback mechanisms in place, and ensure that data quality procedures are maintained, are up-to-date, and are being implemented. For further responsibilities, we refer you to the document [Principles of Data Quality](#) (Chapman 2005a) which should be read as an adjunct to this document.

5. Responsibilities of the Supervisor

The georeferencing supervisor has the principle responsibility for maintaining the quality of the data on a day-to-day basis. Perhaps their key responsibility is to supervise the data-entry procedures (see *Data Entry*, above), and the data validation, checking and cleaning processes. This role is the key role in any georeferencing process, along with that of the data entry operators. It is important that the duties and responsibilities be documented in the institution's best practice manuals and guidelines.

6. Training

Training is a major responsibility of any institution beginning or conducting the georeferencing of their collections. Good training can reduce the level of error, reduce costs and improve data quality. A Georeferencing and Data Cleaning Training Kit is being planned, and hopefully will be developed over the next couple of years. This will aid institutions in training their data entry operators and supervisors in all aspects of the georeferencing and data quality control processes.

7. Performance Criteria

The development of performance criteria is a good way of ensuring a high level of performance, accuracy and quality in the database. Performance criteria can relate to an individual (data entry operator, supervisor, etc.) or to the process as a whole. It can relate to the number of records entered each week, but we would recommend that it relate more to the quality of entry. Where possible, performance criteria should be finite and numeric so that performance against the criteria can be documented. Some examples may include

- 90% of records will undergo validation checking within 6 months of entry,
- any suspect records identified during validation procedures will be checked and corrected within 30 working days,
- feedback from users on errors will be checked and the user notified of the results within two weeks,
- all documentation of validation checks will be completed and up-to-date.

8. Index of Spatial Uncertainty

An Index of Spatial Uncertainty may be developed and documented for the dataset as a whole to allow for overall reporting of the quality of the dataset. This index would supplement a similar index of other data in the database, such as an index of Taxonomic Uncertainty and would generally be for internal use. Currently, no such universal index exists for primary species occurrence data, but institutions may consider developing their own and testing its usefulness. Such indexes should, wherever possible, be generated automatically and produced as part of a data request from the database and packaged with the metadata as part

of the request. Such an index could form the basis for helping users determine the quality of the database for their particular use. The authors of this document would be interested in any feedback from institutions that develop such an index. The index should form an integral part of the metadata for the collection and may include for the georeferencing part of the database:

1. Completeness Index

- percentage of records with georeference fields that have values
- percentage of records with extent fields that have values
- percentage of records with uncertainty fields that have values
- percentage of records with coordinate-precision fields with a value
- percentage of records with datum fields that have a known datum value

2. Uncertainty Index

- average and standard deviation of 'uncertainty' value for those records that have a value
- percentage of records with a maximum uncertainty value in each class
 - a. <100 m
 - b. 100-1,000 m
 - c. 1,000-2,000 m
 - d. 2,000-5,000 m
 - e. 5,000-10,000 m
 - f. >10,000 m
 - g. not determined

3. Currency Index

- time since last data entry
- time since last validation check

4. Validation Index

- percentage of records that have undergone validation test x
- percentage of records that have undergone validation test y, etc.
- percentage of records identified as suspect using validation tests
- percentage of suspect records found to be actual errors

9. Documentation

Documentation is one of the key aspects of any georeferencing process. Documentation involves everything from record-level documentation such as

- how the georeference was determined,
- what method was used to determine the extent and error,
- what modifications were made (for example, if an operator edits a point on the screen and moves it from point 'a' to point 'b' it is best practice to document "why" the point was moved and not just record that location was moved from point 'a' to point 'b' by the operator),
- any validation checks that were carried out, by whom and when,
- flags that may indicate uncertainty, etc.

through to the metadata related to the collection as a whole which may include:

- the overall level of data quality,
-

- the general checks carried out on the whole data set,
- the units of measurement and other standards adopted,
- the guidelines followed,
- the Index of Uncertainty (see earlier discussion, this chapter).

A second set of documentation relates to

- the institution's 'Best Practice' document which we recommend should be derived from this document and tailored to the specific needs of the institution,
- training manuals,
- standard database documentation,
- guidelines and standards.

We recommend that documentation be made an integral part of any georeferencing process.

References

- ADL (Alexandria Digital Library). 2001. *Alexandria Digital Library Gazetteer Server* <<http://www.alexandria.ucsb.edu/gazetteer/>> [Accessed 27 Jan. 2006].
- Bannerman, B.S. 1999. *Positional Accuracy, Error and Uncertainty in Spatial Information*. Geoinnovations, Howard Springs, NT, Australia. <<http://www.geoinnovations.com.au/posacc/patoc.htm>> [Accessed 21 Jan. 2006].
- Chapman, A.D. 2005a. *Principles of Data Quality*. Report for the Global Biodiversity Information Facility 2005. 61pp. Copenhagen: GBIF. <http://circa.gbif.net/Public/irc/gbif/pr/library?!=/webfiles/digit_documents/dataquality_pdf/EN_1.0_&a=d> [Accessed 10 Jan. 2006].
- Chapman, A.D. 2005b. *Principles and Methods of Data Cleaning*. Report for the Global Biodiversity Information Facility 2005. 75pp. Copenhagen: GBIF. <http://circa.gbif.net/Public/irc/gbif/pr/library?!=/webfiles/digit_documents/principlesmethods/EN_1.0_&a=d> [Accessed 10 Jan. 2006].
- Chapman, A.D. 2005c. *Uses of Primary Species-Occurrence Data*. Report for the Global Biodiversity Information Facility 2005. 111pp. Copenhagen: GBIF. <http://circa.gbif.net/Public/irc/gbif/pr/library?!=/webfiles/digit_documents/uses_primary_data/EN_1.0_&a=d> [Accessed 10 Jan. 2006].
- Chapman, A.D., Muñoz, M.E. de S. and Koch, I. 2005. Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context, *Biodiversity Informatics* **2**: 24-41. <<http://jbi.nhm.ku.edu/viewarticle.php?id=9&layout=abstract>> [Accessed 21 Jan 2006].
- Chrisman, N.R. 1991. The Error Component in Spatial Data. pp. 165-174 in: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.
- Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data*. Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens. <http://www.bgbm.org/TDWG/acc/hispid30draft.doc> [Accessed 27 Jan. 2006]
- Conn, B.J. (ed.) 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data*. Version 4 – Internet only version. Sydney: Royal Botanic Gardens. <<http://plantnet.rbgsyd.nsw.gov.au/Hispid4/>> [Accessed 27 Jan. 2006].
- Cullen, A.C. & Frey, H.C. 1999. *Probabilistic Techniques in Exposure Assessment. A Handbook for Dealing with Variability and Uncertainty in Models and Inputs*. New York: Plenum Press. 335 pages.
- DSDM. 2005. *MoSCoW Rules*. Dynamic Systems Development Method Consortium. <<http://na.dsdm.org/en/about/moscow.asp>> [Accessed 10 Jan. 2006].
- Frazier, C., Neville, T., Giermakowski, T. and Racz, G. 2004. *The INRAM Protocol for Georeferencing Biological Museum Specimen Records*. Version 1.3. <http://www.inram.org/modules/UpDownload/store_folder/Documents/INRAM_Biodiversity_Georeferencing_Project/Georeferencing_Guidelines_INRAM-V1.3_2004-03-01.pdf> [Accessed 12 Mar. 2006].
- Geoscience Australia. 2003. *Geodata TOPO-250K (Series 1) Topographic Data* [online]. Geoscience Australia, Canberra. <<http://www.ga.gov.au/nmd/products/digidat/250k.htm>> [Accessed 21 Jan. 2006].
- Geoscience Australia. 2005. *NATMAP Raster Premium. 1:250 000 scale topographic maps of Australia*. On DVD. Canberra: Geoscience Australia.
- Goulet, C.M. 2001. *Magnetic Declinations. Frequently Asked Questions*. Version 4.4. <http://www.geocities.com/magnetic_declination/> [Accessed 7 Jun 2006].
- Gustafson, D.L. and Wefald, M. 2003. *TRS-data*. <<http://www.esg.montana.edu/gl/trs-data>> [Accessed 31 Jan. 2006].
- Hall, B. 1994. *A Review Of The Environmental Resource Mapping System and A Proof That It Is Impossible To Write A General Algorithm For Analysing Interactions Between Organisms Distributed At Locations Described By A Locationally Linked Database and Physical Properties Recorded Within The Database*. Australian Digital Thesis Program: University of Western Sydney. <<http://library.uws.edu.au/adt-NUWS/public/adt-NUWS20040506.151314/index.html>> [Accessed 28 Jan. 2006].

- Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrientes, I. and Rojas, E. 2005. *DIVA-GIS Version 5.2 A geographic information system for the analysis of biodiversity data*. <<http://www.DIVA-GIS.org>> [Accessed 27 Jan. 2006].
- Juran, J.M. 1964. *Managerial Breakthrough*. New York: McGraw-Hill. 396 pp.
- Juran, J.M. 1994. *Managerial Breakthrough*. Rev. edn. New York: McGraw-Hill. 451 pp.
- McElroy, S., Robins, I., Jones, G. and Kinlyside, D. 1998. *Exploring GPS, A GPS Users Guide: The Global Positioning System Consortium*.
- Morton, A. 2006. *UTM Grid Zones of the World*. Digital Mapping Software (DMAP). <<http://www.dmap.co.uk/utmworld.htm>> [Accessed 31 Jan. 2006].
- Museum of Vertebrate Zoology. 2006. *MVZ Guide for Recording Localities in Field Notes*. <http://mvz.berkeley.edu/Locality_Field_Recording_Notebooks.html> [Accessed 24 Jan. 2006].
- National Geophysical Data Center (NGDC). 2004. *Estimated Value of Declination*. <<http://www.ngdc.noaa.gov/seg/geomag/jsp/struts/calcDeclination>>. [Accessed 7 Jun. 2006].
- National Mapping Council of Australia. 1975. *Standards of Map Accuracy*. First Edition February 1953. Second Edition December 1975.
- NIMA, 2000. *Department of Defense World Geodetic System 1984. Its Definition and relationships with local geodetic systems*. TR8350.2, Third Edition, 3 Jan 2000.
- NOAA (2001). *Nautical Charts and DGPS – A warning from NOAA*. <http://www.boatwashington.org/nautical_charts_and_dgps.htm> [Accessed 21 Jan. 2006].
- Srivastava, B.K. and Ramalingam, K. 2006. Error Estimates for WGS-84 and Everest (India-1956) Transformation. *GIS Development* <<http://www.gisdevelopment.net/technology/ip/ma03037abs.htm>> [Accessed 24 Jan. 2006].
- Taylor, C. 2003. *Geographic/UTM Coordinate Converter*. <<http://home.hiwaay.net/~taylorc/toolbox/geography/geoutm.html>> [Accessed 31 Jan. 2006].
- TDWG. 2005). *Geospatial Element Definitions v. 1.4*. Taxonomic Databases Working Group – Extension to Darwin Core Version 2. <<http://darwincore.calacademy.org/Extensions/GeopatialExtension/GeospatialElementDefs>> [Accessed 13 Feb. 2006].
- US Bureau of Budget. 1947. *United States National Map Accuracy Standards*. <<http://rockyweb.cr.usgs.gov/nmpstds/acrodocs/nmas/NMAS647.PDF>> [Accessed 21 Jan. 2006].
- US Navy. *n. dat*. Geographic Datums. PowerPoint from the *Navigator of the Navy* website. <https://www.navigator.navy.mil/navigator/geographic_datums.ppt> [Accessed 24 Jan. 2006].
- Van Sickle, J. 1996. *GPS for Land Surveyors*: Ann Arbor Press, Inc: New York.
- Wheeler, T.A., Huber, J.T. and Currie, D.C. (2001). *Label Data Standards for Terrestrial Arthropods*. Ottawa: Biological Survey of Canada (Terrestrial Arthropods). *Document Series No. 8* <<http://www.biology.ualberta.ca/bsc/briefs/brlabelstandards.htm>> [Accessed 24 April 2006].
- Wieczorek, J. 2001. *MaNIS/HerpNet/ORNIS Georeferencing Guidelines*. University of California, Berkeley: Museum of Vertebrate Zoology. <<http://manisnet.org/GeorefGuide.html>> [Accessed 28 Jan. 2006].
- Wieczorek, J. 2002. *Manual for Georeferencing Calculator*. MaNIS/HerpNet/ORNIS. University of California, Berkeley: Museum of Vertebrate Zoology. <<http://manisnet.org/CoordCalcManual.html>> [Accessed 28 Jan. 2006].
- Wieczorek, J., Q. Guo, and R. Hijmans. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*. 18: 745-767.

Further Reading

- Beaman, R.S. and Conn, B.J. 2003. Automated geoparsing and georeferencing of Malesian collection locality data. *Telopea* 10(1): 43-52.
- Beaman, R.S., Wieczorek, J. and Blum, S. 2004. Determing Space from Place for Natural History Collections. *D-Lib Magazine* 10: 5. <<http://dlib.anu.edu.au/dlib/may04/beaman/05beaman.html>> [Accessed 9 Jun 2006].
- Chapman, A.D. 1999. Quality control and validation of point-sourced environmental resource data pp.409-418 **in** Lowell, K. and Jaton, A. (eds). *Spatial Accuracy Assessment: Land Information Uncertainty in Natural Resources*. Chelsea, Michigan: Ann Arbor Press. 455pp

- Chapman, A.D. 2002. Risk assessment and uncertainty in mapped and modelled distributions of threatened species in Australia pp.31-40 in Hunter, G. & Lowell, K. (eds) *Accuracy 2002 – Proceedings of the 5th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Melbourne: Melbourne University 547 pp.
- FGDC. 2005a. *Proposal for Geospatial Positioning Accuracy Standards, Part 3: National Standard for Spatial Data Accuracy*. Washington, DC: Federal Geographic Data Committee. <<http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part3/progpas3>> [Accessed 5 May 2006].
- FGDC. 2005b. *Proposal for Geospatial Positioning Accuracy Standards*. Washington, DC: Federal Geographic Data Committee. <<http://www.fgdc.gov/standards/projects/FGDC-standards-projects/accuracy/part1/progpas1>> [Accessed 5 May 2006].
- Foote, K.E. and Huebner, D.J. 2000. *The Geographer's Craft Project*. Department of Geography, The University of Colorado at Boulder. http://www.colorado.edu/geography/gcraft/notes/manerror/manerror_f.html [Accessed 17 Jul. 2006].
- Moritz, T. 1999. *Geo-referencing the Natural and Cultural World, Past and Present: Towards Building a Distributed, Peer-Reviewed Gazetteer System*. Presented at the Digital Gazetteer Information Exchange Workshop, Oct 13-14, 1999. Transcribed and edited from audiotape. http://www.alexandria.ucsb.edu/~lhill/dgie/DGIE_website/session1/moritz.htm [Accessed 21 Jan. 2006].
- Stein, B. and Wiczorek, J. 2004. Mammals of the World: MANIS as an example of data integration in a distributed network environment. *Journal of Biodiversity Informatics* 1: 14–22.
- USGS. 2005. *National Mapping program Standards*. Washington: USGS <<http://nationalmap.gov/gio/standards/>> [Accessed 5 May 2006].

Software and on-line Tools

- **BioGeomancer**
<http://www.biogeomancer.org/>
- **BioGeoMancer Classic**
<http://biogeomancer.org/>
- **DIVA-GIS**
<http://www.DIVA-GIS.org/>
- **GeoCalc**
<http://www.geocomp.com.au/geocalc/>
- **GeoLoc – CRIA**
<http://splink.cria.org.br/geoloc?&setlang=en>
- **GEOLocate**
<http://www.museum.tulane.edu/geolocate/>
- **MaNIS Georeferencing Calculator**
<http://manisnet.org/gc.html>
- **NGDC Magnetic Declination Calculator**
<http://www.ngdc.noaa.gov/seg/geomag/jsp/struts/calcDeclination>

Gazetteer Look-up Services

- **Alexander Digital Library Gazetteer Server Client**
<http://middleware.alexandria.ucsb.edu/client/gaz/adl/index.jsp>
- **Fuzzyg – Fuzzy Gazetteer**
<http://tomcat-dmaweb1.jrc.it/fuzzyg/query/>
- **Global Gazetteer**
<http://www.fallingrain.com/world/>
- **GEOnet Names Server**
<http://gnswww.nga.mil/geonames/GNS/index.jsp>

Appendix 1: Guidelines for Georeferencing Locality Types

FEATURE (NAMED PLACE)

Definition:

The simplest locality descriptions consist of only a named place, which is often a feature listed in a standard gazetteer and can probably be located on a map of the appropriate scale.

Feature categories include:

- town, suburb, populated place, or homestead
- spring, bore, tank, well, or waterhole
- island, reef, or cay
- port, bay, gulf, or harbor
- airport, buoy, dock, or jetty
- point, cape, or peninsula
- cave
- dam, or lock
- hill, peak, pass, or mountain
- trig point
- park, reserve, or forestry zone
- junction of two paths (roads, rivers, contour lines, boundaries, etc.)

Despite how they might be presented in a gazetteer or on a map, features are not points; they are areas that have a spatial extent, though the extent may not always be obvious. In a very few cases (such as a trig point), the extent is very small as it is an accurately surveyed point. The important thing is to try to capture the information not only for where the feature is, but also how specific it is (i.e., how big is the extent of the feature).

Some features (e.g., river and road junctions, bridges) may not have gazetteer entries, while others (e.g., properties) may not appear on standard map series. These types of features can be a challenge to locate, and are therefore among the least efficient georeferences to produce. Nevertheless, additional resources, such as internet searches and field notes can often reveal these tricky places

Examples:

Example 1: "Bakersfield"

Example 2: "Point Lookout"

Example 3: "Bennetts Waterhole"

Example 4: "Isla Tiburon"

Example 5: "Lorne Reef"

Example 5: "Yosemite National Park"

Example 6: "Mt Hypipamee"

Example 7: "Junction of Dwight Avenue and Derby Street"

Example 8: "State Forest Reserve 607"

Example 9: "Where Dalby Road crosses Bunya Mountains National Park Boundary"

Example 10: "confluence of Labarge Creek and South Labarge Creek"

Example 11: "At 100 m contour line on Black street"

Example 12: "junction of Rio Claro and Rio La Hondura"

Example 13: "Victoria River Station" [Northern Territory, Australia]

Georeferencing Procedure:

Features with an obvious spatial extent — use the geographic center (i.e., the midpoint of the extremes of latitude and longitude) for the coordinates. If the geographic center does not fall inside the shape of the shaded area, then pick the nearest point to the center that lies within the shape (see Figure 4). Use the distance from the coordinates to the furthest point within the named place as the extent. Some gazetteers give bounding boxes to describe the extents

of large places and you can use these to determine the extent by measuring them from a map or by using a geographic distance calculator such as the [Perpendicular Distance Calculator](#)¹⁵ from the Center for Biodiversity and Conservation (CBC).

Features without an obvious spatial extent — some features do not have a shaded boundary or a topographic symbol for buildings shown on the map (especially for non-USA locales). Some of these features may have large, but indistinct extents (mountains, trap lines). Other features may be relatively small (springs, junctions), with no apparent extent on a map. Use and document your judgment when placing the coordinates and estimating the extent of large features, and use a standard extent for small features based on the feature type. The extent of road junctions, for example, cannot be measured on maps, so use the following extent recommendations from Frazier *et al.* (2004):

- For 2-lane city streets and 2-lane highways, the extent is 10 m.
- For 4-lane highways, the extent is 20 m.
- For large highways with medians, the extent is 30 m.
- If unknown, use 15 m.

It is worthwhile to create a feature type extent table as part of your institutional best practices document so that there is consistency in extents for features whose size cannot be measured without ground-truthing.

'Exact' locations — if the locality appears to be 'exactly at' the locality cited (GPS reading) use the accuracy of the GPS as the extent.

In some cases – for example, an accurately recorded trig point – the extent and the uncertainty may be identical, however, collections are seldom made at the exact locality cited (e.g., right on top of the trig point), so the extent is usually much larger than a literal reading might suggest.

If you choose to use a gazetteer to obtain coordinates, keep in mind that they may not be at the geographic center of the feature. For example, the coordinates of a populated place may be at the main post office or the courthouse (if that place is a county seat). Coordinates for rivers and streams are usually at the mouth. For this reason, it is a good idea to use the gazetteer coordinates to find the feature on a map, and then use the map to find the geographic center of the feature.

When recording the method of determination of the coordinates and uncertainty in the remarks, use "measured from the main post office" or "measured from the geographic center of Bakersfield", etc.

Care! Some older gazetteers reference the bottom left hand corner of the position where the name is to appear on a printed map rather than the actual location of the feature. Most gazetteers have been fixed in recent years, but care should be taken when using an unfamiliar gazetteer. Always check the map, which you will need to do in any case to calculate the extent.

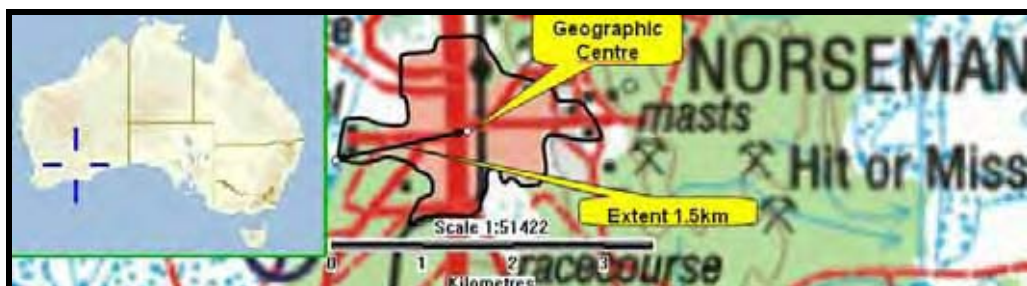


Fig. 4. Calculation of the geographic center and extent for Norseman, Western Australia. Background map from Geosciences Australia (2005).

¹⁵ Perpendicular Distance Calculator <http://geospatial.amnh.org/open_source/pdc/index.html>.

Subdivisions of a feature — such as “N part of Mono Lake” calculate the extent based only on the subdivision and proceed as you would with a location consisting of a named place with a spatial extent.

Properties (ranches, farms, stations, etc.) — if you are unable to locate them in gazetteers or on regular maps, you may have to use a cadastral map, or carry out a search to see if you can locate them in relation to nearby cities or other geographic entities. If you are unable to find the boundaries, and thus determine the geographic center, then use the coordinates of the homestead or major property buildings and estimate the size of the property from the location of other buildings not on that property.

Caves — use the coordinates of the entry to the cave. This will usually be the location given in a gazetteer or on a regular map.

Uncertainty:

Use the *MaNIS Georeferencing Calculator* (<http://manisnet.org/gc.html>) to determine “Maximum Uncertainty Distance”.

- For **Calculation Type** use:
“Error – enter Lat/Long for the actual locality”
- For **Locality Type** use:
“Named place only”.

See Example 1, below

Example 1.

Locality: “Bakersfield”

Suppose the coordinates for Bakersfield came from the GNIS database (a gazetteer) and the distance from the center of Bakersfield to the furthest city limit is 3 km.

Coordinate System: degrees minutes seconds

Latitude: 35° 22' 24" N

Longitude: 119 ° 1' 4" W

Datum: not recorded; 79 m uncertainty

Coordinate Precision: nearest second; 40 m uncertainty

Coordinate Source: gazetteer

Extent of Named Place: 3 km

Distance Units: km

Decimal Latitude: 35.37333

Decimal Longitude: -119.01778

Maximum Uncertainty Distance: 3.119 km

NEAR A FEATURE

Definition:

A locality given without an exact position, but with “near”, “in the vicinity of”, “adjacent to”, or some similar relation to a feature cited.

These locality descriptions imply an offset from a named place without definitive directions or distances.

Examples:

Example 1: “Near Las Vegas“

Example 2: “vicinity of Tumberumba“

Example 3: “Big Bay vicinity”

Example 4: “near MS 117 on Dalton Hwy”

Example 5: “near Bend 43 on Great Western Hwy”

Example 6: “vicinity of bridge over Condamine River on Warrego Highway“

Example 7: “adjacent to railway underpass on Smith Street”

Example 8: “area of confluence of Black and Oshetna Rivers”

Georeferencing Procedure:

In these cases use the geographic center of the named place for the geographic coordinates.

If you are unable to determine the exact coordinates of the locality, then use the coordinates as near as possible to the referenced locality (and on the path if appropriate).

Extent:

The extent should be calculated as the greater of 2km or 200% of the extent of the named place. Clearly there is a measure of subjectivity involved here and you should use your judgement and evidence from other sources. Let common sense prevail and document the assumptions made.

Uncertainty:

Calculate the same as for '**Feature**' but note the increase in extent.

BETWEEN TWO FEATURES

Definition:

A locality cited as 'between' two features or named places.

Examples:

Example 1: "between Point Reyes and Inverness"

Image

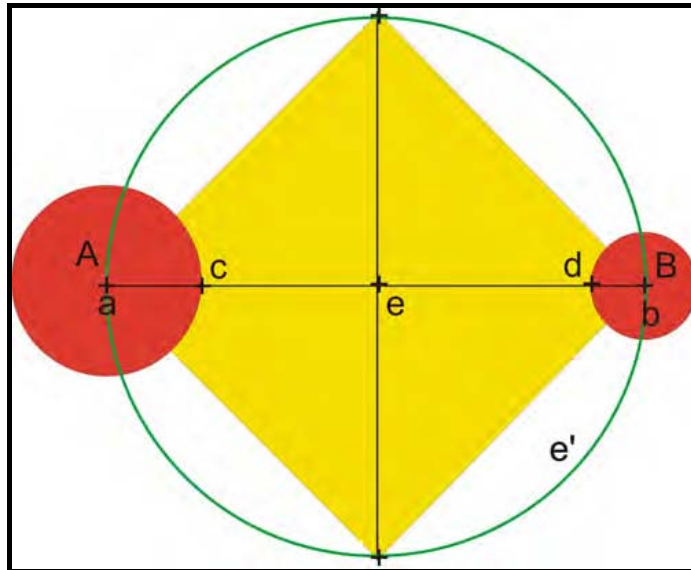


Fig. 5. The diagram above illustrates the general case of a locality description of the type "Between A and B."

Georeferencing Procedure:

Find the coordinates of the midpoint between the centers of the two named places (the point **e** in Figure 5).

Extent:

Use one-half the distance between the centers of A and B.

Uncertainty:

Calculate the same as for '**Feature**'.

STREET ADDRESS

Definition:

Locality is a street address – usually with a number, a street name, and a feature name.

In some places, street numbers in rural areas represent a metric distance from the start of the road.

Examples:

Example 1: “1 Orchard Lane, Berkeley, CA”

Example 2: “21054 Baldersleigh Road, Guyra, NSW” (indicates that the locality is 21.054 km from the beginning of Baldersleigh Road).

Example 3: “Backyard of 593 West Street, Louisville, Boulder County, Colorado”

Example 4: “Greenhouse at 20th and Broadway, Boulder, Boulder County, Colorado”

Georeferencing Procedure:

Addresses are sometimes given when specimens are collected in cities or towns. When possible, plot the point at the indicated spot with the aid of a local road map or a mapping product such as [Google Maps®](#), [Maporama®](#) or [Mapquest®](#), White or Yellow Pages directory, or a GPS. If the exact address cannot be found, estimate the location as well as possible. Remember that many addresses reflect a grid system of labeling addresses. For instance, addresses between 12th Street and 13th Street would lie between 1200 and 1300. Be aware, however, that street names often change over time. To your best ability, locate this area on the map or map software you are using to get the coordinates, if the electronic gazetteer does not display them automatically.

Building names are often given to clarify the location within a town or city. Rarely are these buildings given coordinates in a gazetteer; however, they can sometimes be located using a Yellow Pages directory, which may be available via the Internet. Unlike natural features, most buildings change names or even disappear over time, so verify that the building named in the record existed in that location at that time.

Extent:

Use as the extent the smallest area that is identifiable and that cannot be any other address. If you cannot determine the location and size of an address within a block, use half the length of the city block for the extent and make note of this in the georeferencing remarks.

Uncertainty:

Calculate the same as for '**Feature**'.

PATH

Definition:

The locality is a linear feature such as a road, trail, boundary, river, or contour line. The locality may also refer to part (or subdivision) of the path (see Examples 5-7). Localities that are given without an exact position, but are cited as “near”, “in the vicinity of”, “adjacent to”, a path such as a road, river, etc. (see Examples 8 and 9) are treated in the same manner as any other path, but with perhaps a wider footprint – see also under ‘**Near a Feature**’, above.

Note! A path clause in a locality description is often meant to be read in combination with another clause. The relationship between the path clause and other clauses in the same description is important, because the resulting shape will be affected. For example, a description with a path followed by an offset from a feature at a heading (“Hwy 101, 2 mi N Santa Rosa) should actually be calculated as a clause of the type “offset from a feature at a heading along a path” rather than as the intersection of a path and a clause of the type “offset from a feature at a heading”

Examples:

Example 1: “Hwy 1”

Example 2: “Nepean River”

Example 3: “along 100 m contour line”

Example 4: “N. Boulder Creek, 1.3 miles above Boulder Falls”

Example 5: “mouth of Goodpaster River”

Example 6: “head of Mooney Creek”

Example 7: “Eastern part of Logan Motorway”

Example 8: “vicinity of Uyamitquaq Ck.”

Example 9: “adjacent to eastern boundary of Foz do Iguaçu Park”

Georeferencing Procedure:

Roads — a path may be defined with reference to a named place (see Note above). This may influence where one places the coordinates. If there is no further refinement, then treat the road similarly to a river, as explained below. If there is reference to an offset or a position on the path, then treat the location as any other feature and refer to the appropriate sections, such as ‘**Offset**’, or ‘**Feature**’.

Rivers — if you are unable to traverse the length of the river to find the geographic center, then make a straight line from the mouth of the river to the head of the river (or the extreme points within the county, state, etc. you are concerned with). Find the center of this line, and place your coordinate point closest to the center of the line on the river itself (see Figure 6). This method may lead to large errors in rivers that have large changes in direction. Use your common sense to determine the most appropriate point, bearing in mind the suggested methods above.

The mouth of a river is not always easy to determine, but is usually taken to be formed by a straight line across the river at the position where the river joins a larger body of water (sea, bay, lake, another river, etc.). In some rare cases, it may refer to the downstream end where the river changes its name. It is the position of lowest elevation of the river.

Similarly, the head of a river (where the river begins) can also be difficult to determine. Though the head is always the highest point on the part of the river bearing the same name, it may begin in a mountain, canyon, or lake, and may need to be estimated because it has become too fine or broken up into smaller streams to accurately identify on a map.

Sometimes the terms ‘above’, ‘below’, ‘left bank’, or ‘right bank’ are used with rivers instead of cardinal directions (see Example 4, above). ‘Above’ is used when referring to upstream of the feature while ‘below’ refers to downstream. The direction a river flows can be easily determined on a topographic map by looking at the contour lines and elevation. The contour

lines will always point upstream as they cross the river. The terms left and right bank refer to the side of the river when facing downstream.

'Mouth of River' (Example 5) and 'head of River' (Example 6) are usually best treated as you would a **'Feature'**.

Care! When using older maps, be aware that rivers may have changed course and may have been in a different location at the time the collection was made, compared with the position drawn on the map at a different time. In addition, the apparent position of the mouth of a river can be strongly influenced by the scale of the map being used.

Note! Do not use the coordinates given by gazetteers, as these points usually correspond to the mouth of the river, not the geographic center.

Contour Lines — If the contour line has ends within the area of interest, treat it the same as you would a river. If the contour line is closed (i.e., forms a polygon around a hill or mountain, etc.), then treat the enclosed area the same as you would a **'Feature'** and use the geographic center of the polygon for the geographic coordinates.

Subdivisions of a Path — where a subdivision of a path still describes a path, continue to treat it as a road, river or contour line as above. In Example 7, for instance, you may take the midway point on the Logan Motorway as the western limit of the subdivision meant by the 'eastern part'. Use that limit as the basis to determine both the coordinates and the extent.

Image:

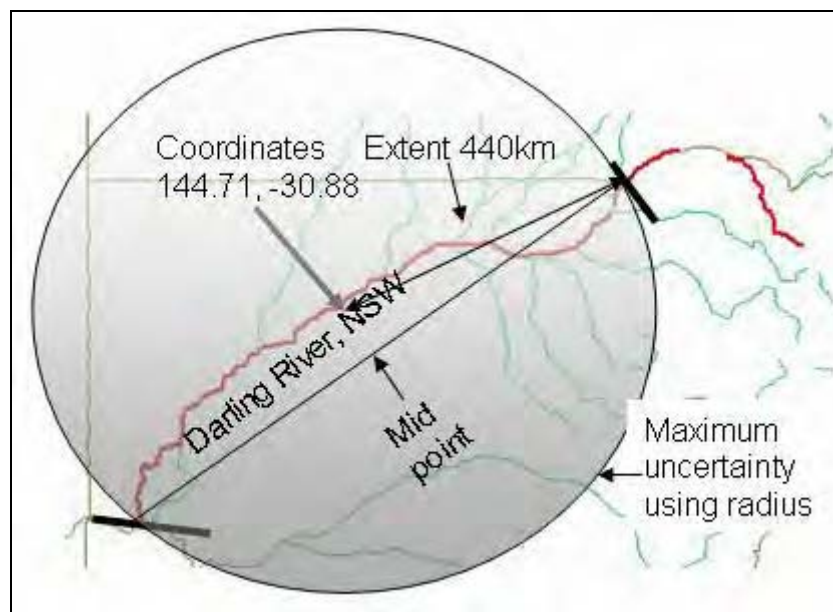


Fig. 6. An example of determining coordinates and extent for a path (in this case the Darling River in New South Wales, Australia). Use of an uncertainty polygon would give a more accurate representation for maximum uncertainty than the point-radius method.

Extent:

The extent is the distance from the point nearest the geographic center of the path to the point on the path furthest from that center point. Make sure to base the center point on only that portion of the path within the boundaries of interest.

Uncertainty:

Calculate the same as for **'Feature'**.

BETWEEN TWO PATHS

Definition:

A locality cited as being between two paths (two roads, two rivers, a road and a river, etc.).

Examples:

Example 1: “between Tanama R. and Clearwater Ck.”

Example 2: “between Aldersley and Bridge Streets” (i.e., two streets that don’t intersect)

Example 3: “on Hwy 14, between highway and adjacent fence”

Georeferencing Procedure:

Create a polygon from the two paths and the end points of each of the paths – for example, the state boundary, where the river joins another river or changes names, a road intersection, etc. (see Figure 7.)

Once the polygon is drawn – then the coordinates are determined in the same manner as for a *‘Feature’*, above.

Image:

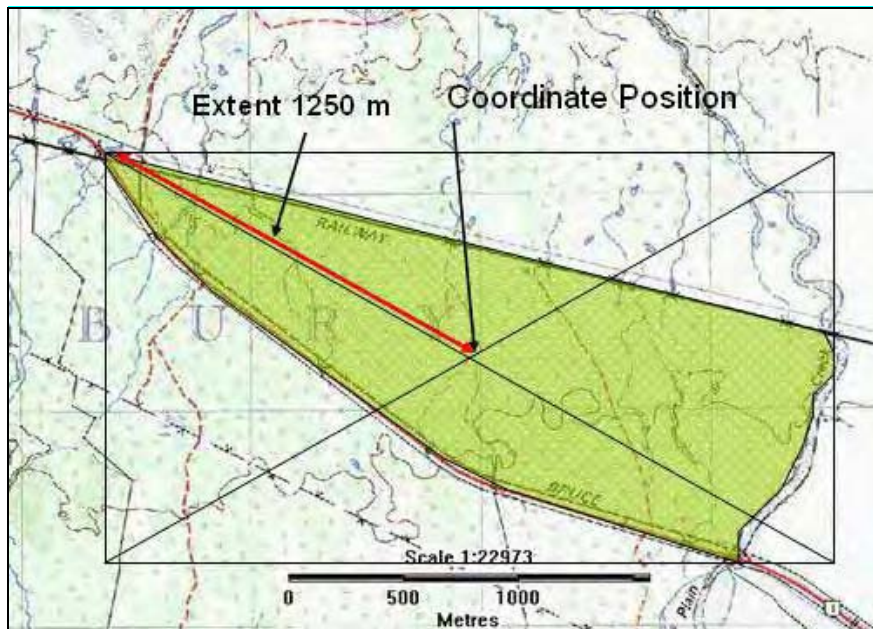


Fig. 7. An example of determining coordinates and extent for a location between two paths (in this case “between the Bruce Highway and the Railway line, West of Plain Creek and before the Railway crossing on the Highway”). Background map from Geosciences Australia (2005).

Extent:

Once the polygon has been drawn as above, then the extent is determined in the same way as for a *‘Feature’*.

Uncertainty:

Calculate the same as for *‘Feature’*.

OFFSET DISTANCE

Definition:

Locality consists of an offset from a named place without any direction specified.

Offsets without a direction are often the result of errors by the collector when recording the locality. Occasionally, these localities are data entry errors. Try to view the original collection catalogs or labels, as there may be more information in them.

Examples:

Example 1: "5 km outside Calgary"

Example 2: "15 km from Recife"

Image:

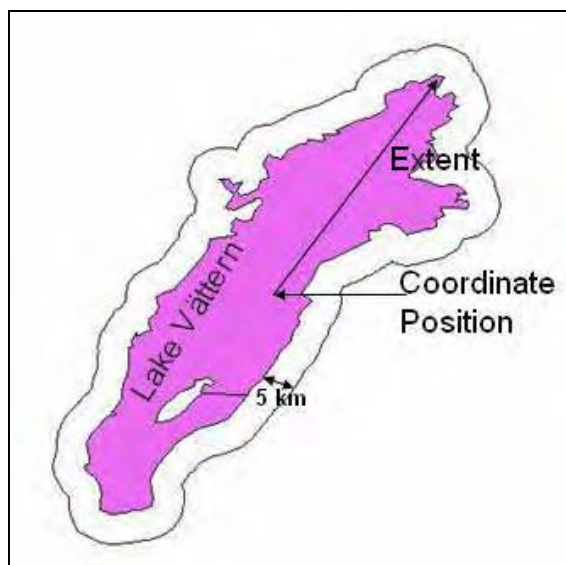


Fig. 8. An example of determining coordinates and extent for a location with offset distance only (in this case "5 km from Lake Vättern, Sweden"). The coordinates are 14.56° , 58.30° . Offset distance is 5 km, extent of the named place (Lake Vättern) is 61.2 km. These values are then used in the MaNIS Georeferencing Calculator to determine maximum uncertainty.

Georeferencing Procedure:

Record the geographic coordinates of the center of the named place, as you would for a normal '**Feature**'.

Sometimes offset information is vague either in its direction or in its distance. If the direction information is vague, record the geographic coordinates of the center of the named place and include the offset distance in the determination of the maximum uncertainty (see figure 8).

Extent:

Use the extent of the named place.

Uncertainty:

Use the **MaNIS Georeferencing Calculator** (<http://manisnet.org/gc.html>) to determine "Maximum Uncertainty Distance".

- For **Calculation Type** use "Error – enter Lat/Long for the actual locality"

- For **Locality Type** use “Distance only (e.g., 5 mi from Bakersfield)”.

Example 1.

Locality: “5 mi from Bakersfield”

Suppose the coordinates for Bakersfield came from Topozone® with the map coordinates reprojected in NAD27. Suppose also that the distance from the center of Bakersfield to the furthest city limit is 2 mi.

Coordinate System: decimal degrees
Latitude: 35.373
Longitude: -119.018
Datum: NAD27; no uncertainty
Coordinate Precision: 0.001 degrees; 0.089 mi uncertainty
Coordinate Source: gazetteer
Offset Distance: 5 mi
Extent of Named Place: 2 mi
Distance Units: mi
Decimal Latitude: 35.373
Decimal Longitude: -119.018
Maximum Uncertainty Distance: 8.089 mi

Example 2.

Locality: “5 km from Lake Vättern, Sweden” (see Figure 8).

Coordinate System: decimal degrees
Latitude: 58.30
Longitude: 14.56
Datum: unknown
Coordinate Precision: 0.001 degrees, 1520 m uncertainty
Coordinate Source: gazetteer
Offset Distance: 5 km
Extent of Named Place: 61.2 km
Distance Units: km
Decimal Latitude: 58.30
Decimal Longitude: 14.56
Maximum Uncertainty Distance: 68.559 km

NB! See discussion on “Estimating Uncertainty from Distance” earlier in this document. The Georeferencing Calculator uses the method for estimating uncertainty given in Wieczorek *et al.* (2004).

OFFSET DIRECTION

Definition:

Locality consists of a direction from a named place without any distance specified.

Examples:

Example 1: "N Palmetto"

Example 2: "W of Jondaryan"

Image:

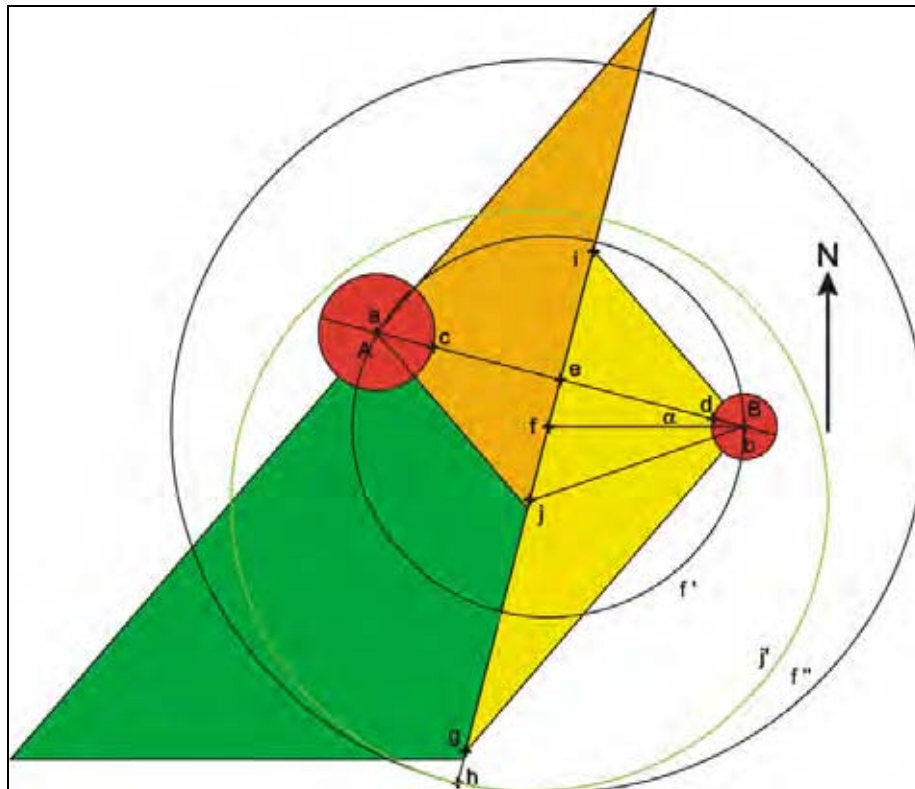


Fig. 9. An example of a locality description of the type "At a heading from B". In this diagram the specific example is "West of B". The area corresponding to "West of B" is encompassed by the bright yellow triangle connecting the three points b, i, and g. The orange triangle would be interpreted as "East of A" and the green triangle would be interpreted as "South of A".

There are a number of ways one might calculate the coordinates, extent, and uncertainty for this complicated scenario, some of which are described below using the values shown in Figure 9.

Alternative 1:

Coordinates: Place the coordinates at the point **f**, which is at a distance $r/\cos(\alpha)$ W from the center of B, where r is one half the distance between the centers of A and B and α is the angle between west and the direction from the center of B to the center of A.

Extent: The radius of **f'**, which is $r/\cos(\alpha)$.

Disadvantage: This alternative leaves out some of the triangle (**big**).

Advantages:

1. The center of the uncertainty radius is the point due W furthest from the center of B within the triangle (**big**).
2. This is the simplest of the three alternatives to calculate.

Alternative 2:

Coordinates: Place the coordinates at the point **f**, as in Alternative 1.

Extent: The radius of **f''**, which is the extent of B plus $r*\sqrt{2}/(2*\cos(\alpha)*\sin(\theta-\alpha))$, where the angle θ is based on the direction uncertainty (45 degrees for West).

Disadvantage: The radius of uncertainty is larger than it needs to be to cover the area that might reasonably be called 'West of B'.

Advantage:

1. This alternative leaves out none of the triangle (**big**).
2. It has its center on the point furthest due W of the center of B within the triangle (**big**).

Alternative 3:

Coordinates: Place the coordinates at the point **j**, which is half-way between the points **i** and **g**. The coordinates for this point is beyond the ability of a georeferencer to discern.

Extent: The radius of **j'**, which is the extent of B plus $r*(\tan^2(\theta-\alpha)+1)/(2*\tan(\theta-\alpha))$, where θ is the direction uncertainty (45 degrees for West).

Disadvantage:

1. This alternative leaves out none of the triangle (**big**).
2. The coordinates of the point **j** cannot be determined readily from a map; they have to be calculated.
3. The uncertainty for this alternative is the most complex to calculate.
4. The center of the uncertainty is not due W of the center of B.

Advantage:

1. This alternative leaves out none of the triangle (**big**).
2. The size of the uncertainty radius is as small as it can be and still encompass the whole triangle (**big**).

Alternative 4:

NoGeorefBecause: "no offset distance given".

Disadvantage:

1. No georeference is produced.

Advantage:

1. This alternative avoids all of the subjectivity required to interpret this vague description.

Georeferencing Procedure:

When only an offset is given with no distance, it is virtually impossible to georeference with certainty without additional information. For example, if we have a location 'East of Albuquerque' with no other information, there is no clear indication of how far one should go 'East' to find the location – to the next nearest named place; the next nearest named place of equivalent size, or just keep going? In reality, such a description could describe half of the Earth's surface. It is for this reason that we recommend using Alternative 4 above.

Seldom is such information given alone; there is usually some supporting information. For example, the locality may have higher-level geographic information such as 'East of Albuquerque, Bernalillo County, New Mexico'. This gives you a stopping point (the county), and should allow you to georeference the locality.

Uncertainty:

The *MaNIS Georeferencing Calculator* (<http://manisnet.org/gc.html>) does not explicitly calculate coordinates and uncertainty for this locality type. Nevertheless, the uncertainty can be calculated for any of the first three alternatives given above if one first determines the coordinates and extent. The Georeferencing Calculator can be used in two steps to georeference using Alternative 1, above. For the first step, determine the coordinates for the point **f**:

- For **Calculation Type** use
"Coordinates and error"
- For **Locality Type** use
"Distance at a heading".



Fig. 10. Extract from TOPO250K digital map showing Jondaryan, Queensland, Australia. Map from Geosciences Australia (2005).

Example 1. (see Figure 10)**Locality: "W of Jondaryan"**

Suppose the coordinates for Jondaryan came from a gazetteer using the Australian Geodetic Datum 1984 (AGD84). Malu is the next populated place in a westerly direction from Jondaryan at a distance of 3.65 and a heading of 305°. The scale of the map is 1:250,000 and the metadata for the map indicates an uncertainty of ~160 m (see Table 5).

Coordinate System: degrees, minutes, seconds

Latitude: 27° 21' 50" S (for Jondaryan)

Longitude: 151° 34' 59" E (for Jondaryan)

Datum: AGD 84; no uncertainty

Coordinate Precision: 1 second

Offset distance: 4.46 km ($r/\cos(\alpha)$ where $r = 3.65$ km and α is the difference between the heading 305° and west 270°, or 35°).

Direction: W

Decimal Latitude: -27.36389 (for point **f**)

Decimal Longitude: 151.53797 (for point **f**)

Maximum Uncertainty Distance: 1.592 km

For the second step, determine the maximum uncertainty distance from the point **f**:

- For **Calculation Type** use
"Error – enter Lat/Long for the actual locality"
- For **Locality Type** use
"Named place only".

Example 1. Step 2.

Coordinate System: degrees, minutes, seconds

Latitude: 27° 21' 50" S

Longitude: 151° 32' 16.69" E (based on the Decimal Longitude from Step 1)

Datum: AGD 84; no uncertainty

Coordinate Precision: 1 second

Coordinate Source: non-USGS map: 1:250,000; 0.16 km uncertainty.

Extent of Named Place: 4.46 km (radius of **f** in Figure 9)

Distance Units: km

Distance Precision: 1/100 km

Decimal Latitude: -27.36389

Decimal Longitude: 151.53797

Maximum Uncertainty Distance: 4.751 km

OFFSET AT A HEADING

Definition:

The locality contains a distance in a given direction from a feature or named place. There are several variations on such localities.

Localities that have one linear offset measurement from a named place, but do not specify how that measurement was taken (see Example 1, below), are open for case-by-case judgment. The judgment itself must be documented in the remarks for the determination (e.g., 'Assumed "by air" – no roads E out of Yuma', or 'Assumed "by road" on Hwy. 80'). In this case, the remark should be something like 'Uncertainty encompasses both distance by air and distance by road on Hwy. 80').

In Example 2, the locality is on the east side of the river, in Illinois, rather than on the west side, in Missouri. In this example, the 16 miles were assumed to be 'by air' – but see similar example under in the next Locality Type: **Offset along a Path**.

The addition of an adverbial modifier to the distance part of a locality description, while an honest observation, should not affect the determination of the geographic coordinates or the maximum uncertainty. In Example 3, below, treat the locality as if it read "25 km WNW of Campinas"

Examples:

Example 1: "10.2 mi E of Yuma"

Example 2: "16 mi from St Louis on left bank of the Mississippi River – downstream"

Example 3: "about 25 km WNW of Campinas"

Example 4: "10 mi E (by air) Bakersfield"

Georeferencing Procedure:

Use the geographic coordinates of the named place (see '**Feature**', above) as a starting point. Sometimes the locality description gives a method for determining the offset (e.g., 'by road', 'by river', 'by air', 'up the valley', etc.) For all cases except 'by air' (see Example 4), use the next Locality Type: '**Offset along a Path**', below.

Where the method of determining the offset cannot be determined from the locality description or additional information and there is no obvious major path that can be followed in the rough direction and distance given, assume the collector measured the distance by air.

If there is no clear best choice between 'by road' or 'by air', you may wish to use the midpoint between the two possibilities as the geographic coordinate and assign an uncertainty large enough to encompass the coordinates and uncertainties of both methods. This choice is not recommended here for two reasons. First, the resulting coordinates will not match either of the two possible interpretations. Second, it will take about three times as long to calculate since the two interpretations have to be made, followed by the determination that encompasses them both. Since the offset at a heading "by air" will usually encompass the alternative by road anyway, this is the recommended option. You can increase the maximum uncertainty to encompass the other possible choice. Once again, this recommendation applies if you don't have a compelling reason to use the offset along a path.

To calculate the coordinates, use the geographic coordinates of the center of the named place as a starting point (in the Example 1 above, use the center of Yuma) and enter its coordinates and extent in the [MaNIS Georeferencing Calculator](#) using the Calculation type: '**Coordinates and Error**'. Enter the distance and direction given – (make sure relevant parameters are filled in or selected, such as datum, direction, offset distance, distance units and precision, and coordinate source, system, and precision,) and push "Calculate." The new coordinates that appear at the bottom of the calculator are the ones you can now enter in your database. They should be different from the coordinates you entered in the 'Latitude' and 'Longitude' spaces – if they are not, check to make sure you have chosen the correct

Calculation Type. You should also check the resulting locality coordinates on a map (or for the USA, in Topozone.com) to make sure they make sense. Be sure to choose the same datum as the original coordinates when viewing the result.

Extent:

The extent is the extent of your starting point - usually a named place such as a city or a junction.

Uncertainty:

Use the *MaNIS Georeferencing Calculator* (<http://manisnet.org/gc.html>) to determine "Maximum Uncertainty Distance".

- For **Calculation Type** use "Coordinates and error"
- For **Locality Type** use "Distance at a Heading".

Example 1.

Locality: "10 mi E (by air) Bakersfield"

Suppose the coordinates for Bakersfield came from the GNIS database (a gazetteer), the coordinates of the locality were calculated to the nearest second, and the distance from the center of Bakersfield to the furthest city limit is 2 mi.

Coordinate System: degrees, minutes, seconds

Latitude: 35° 22' 24" N

Longitude: 118° 50' 56" W

Datum: not recorded; 0.049 mi uncertainty

Coordinate Precision: nearest second; 0.024 mi uncertainty

Coordinate Source: gazetteer (for which the datum is unknown)

Offset Distance: 10 mi

Extent of Named Place: 2 mi

Distance Units: mi

Distance Precision: 1 mi

Direction Precision: E (45 degrees precision – between NE and SE)

Decimal Latitude: 35.37333

Decimal Longitude: -118.67179

Maximum Uncertainty Distance: 16.588 mi

Example 2.

Locality: "10 mi ENE (by air) Bakersfield"

Suppose the coordinates for the locality were interpolated to the nearest second from the USGS Gosford 1:24,000 Quad map and the distance from the center of Bakersfield to the furthest city limit is 2 mi.

Coordinate System: degrees, minutes, seconds

Latitude: 35° 24' 21" N

Longitude: 118° 51' 25" W

Datum: NAD27; no uncertainty

Coordinate Precision: nearest second; 0.024 mi uncertainty

Coordinate Source: USGS map: 1:24,000; 0.008 mi uncertainty

Offset Distance: 10 mi

Extent of Named Place: 2 mi

Distance Units: mi

Distance Precision: 10 mi

Direction Precision: ENE (11.25 degrees either side of ENE)

Decimal Latitude: 35.46134

Decimal Longitude: -118.69326

Maximum Uncertainty Distance: 12.379 mi

OFFSET ALONG A PATH

Definition:

Locality describes a route from a named place.

If the distance was along a linear feature such as a road or river, measure along the feature for the distance and in the direction cited, rather than use a straight line. There is no uncertainty due to direction imprecision.

Examples:

Example 1: "7.9 mi N Beatty, on US 95"

Example 2: "13 mi E (by road) from Bakersfield"

Example 3: "18 km W of Guyra, on Baldersleigh Road"

Example 4: "2km downstream from Wallaman Falls"

Example 5: "3 km above Anita Grande on Rio Jimenez"

Example 6: "left bank of the Mississippi River, 16 mi downstream from St. Louis"

Georeferencing Procedure:

If 'by road' is specified in the locality description, or if there is an obvious major road that can be followed that complies to the direction and distance exactly, you can assume that the collector traveled by road. If there is a choice between multiple roads that fit the description, choose one of them as the basis for the georeference and increase the error to encompass the other possible choices.

Use the center of the starting point (in the Example 1, above, use the center of Beatty), and use the measuring tool found in [Terrain Navigator](#)¹⁶ (USA only), or your own appropriate application, to follow the road until you have gone the distance cited. Take the coordinates from this ending point. Be sure to make a note of the name of the road you followed in the Remarks if it is not already in the locality description.

Extent:

The extent is the extent of your starting point - usually a named place such as a city or a junction.

Uncertainty:

Use the *MaNIS Georeferencing Calculator* (<http://manisnet.org/gc.html>) to determine "Maximum Uncertainty Distance".

- For **Calculation Type:** use
"Error – enter Lat/Long for the actual locality"
- For **Locality Type:** use
"Distance along a Path".

¹⁶ Terrain Navigator®, <<http://www.maptech.com/land/index.cfm>>.

Example 1.**Locality: "13 mi E (by road) Bakersfield"**

Suppose the coordinates for this locality were interpolated to the nearest 1/10th minute from the USGS Taft 1:100,000 Quad map and the distance from the center of Bakersfield to the furthest city limit is 2 mi.

Coordinate System: degrees, decimal minutes

Latitude: 35° 26.1' N

Longitude: 118° 48.1' W

Datum: NAD27; no uncertainty

Coordinate Precision: 0.1 minutes; 0.148 mi uncertainty

Coordinate Source: USGS map: 1:100,000; 0.032 mi uncertainty

Extent of Named Place: 2 mi

Distance Units: mi

Distance Precision: 1 mi

Decimal Latitude: 35.43500

Decimal Longitude: -118.80167

Maximum Uncertainty Distance: 3.180 mi

OFFSET IN ORTHOGONAL DIRECTIONS

Definition:

Locality consists of a linear distance in each of two orthogonal directions from a named place (Figures 11 and 12).

Examples:

Example 1: "2 mi E and 1.5 mi N of Bakersfield"

Example 2: "6 km N and 4 km W of Welna"

Example 3: "2 miles north, 1 mile east of Boulder Falls, Boulder County, Colorado"

Georeferencing Procedure:

Where localities have two orthogonal measurements in them, it should always be assumed that the measurements are 'by air' unless there is a reference that indicates otherwise.

Use the center of the starting point (e.g., in the Example 2, above, use the center of Welna), and enter its coordinates and extent in the [MaNIS Georeferencing Calculator](#) using the Calculation Type: "Coordinates and Error" Enter the distances and directions given, and push "Calculate." The new coordinates that appear at the bottom of the calculator are the ones you can now enter in your database. They should be different from the coordinates you entered in the 'Latitude' and 'Longitude' spaces – if they are not, check to make sure you have chosen the correct Calculation Type.

Figures:

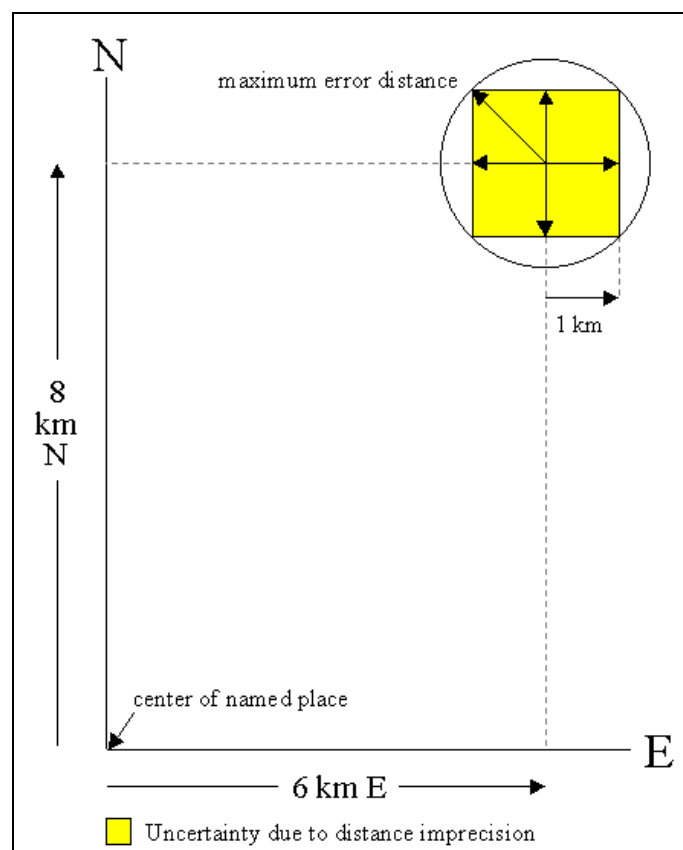


Fig. 11. Example of calculating maximum uncertainty using distance imprecision for two orthogonal offsets from the center of a named place. From Wieczorek (2001).

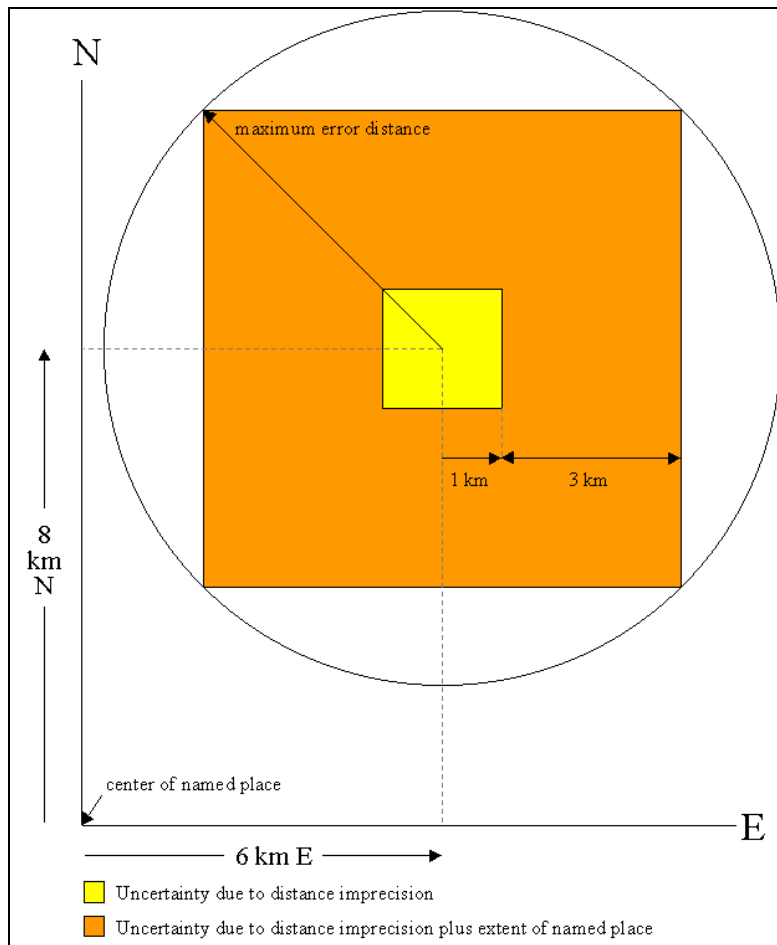


Fig. 12. Calculating maximum uncertainty from the combination of distance imprecision and extent. From Wieczorek (2001).

Extent:

Extent is the extent of your starting point - usually a city or a junction.

Uncertainty:

Use the **MaNIS Georeferencing Calculator** (<http://manisnet.org/gc.html>) to determine "Maximum Uncertainty Distance".

- For **Calculation Type** use "Coordinates and error"
- For **Locality Type** use "Distance along Orthogonal Directions".

Example 1.**Locality: "2 mi E and 3 mi N of Bakersfield"**

Suppose the coordinates for Bakersfield (the named place) came from the GNIS database (a gazetteer), the coordinates of the locality given to the nearest second, and the distance from the center of Bakersfield to the furthest city limit is 2 mi.

Coordinate System: degrees, minutes, seconds

Latitude: 35° 25' 4" N

Longitude: 118° 58' 54" W

Datum: not recorded; 0.049 mi uncertainty

Coordinate Precision: nearest second; 0.024 mi uncertainty

Coordinate Source: gazetteer

North or South Offset Distance: 3 mi

North or South Offset Direction: N

East or West Offset Distance: 2 mi

East or West Offset Direction: E

Extent of Named Place: 2 mi

Distance Units: mi

Distance Precision: 1 mi

Decimal Latitude: 35.4621

Decimal Longitude: -118.94623

Maximum Uncertainty Distance: 4.337 mi

OFFSET FROM TWO DISTINCT PATHS

Definition:

Locality consists of orthogonal offset distances, one from each of two different paths. This is a more unusual situation, but it does occur.

Examples:

Example 1: "1.5 mi E of LA Hwy. 1026 and 2 mi S of U.S. 190"

Georeferencing Procedure:

Locating the coordinates of a position like this is tricky. To do so, you have to draw a path parallel to, and at the appropriate distance and heading from, each of the reference paths. The place where they intersect (hopefully there is only one) is the coordinate.

Extent:

Use the extent of the wider of the two paths from which you are measuring. The uncertainty from the width of the wider highway will completely encompass the uncertainty from the narrower one. In Example 1, above, Interstate 190 is a big four-lane highway and LA 1026 is a two-lane highway. Since the measurements are orthogonal to each other and to the roads in this case, each extent would be half of the width of the respective highways. Since I 190 is the larger of the two, its extent would completely encompass the extent from LA 1026.

For standard extents of roads, use the values described under Locality Type: **Feature**, above.

Uncertainty:

Use the **MaNIS Georeferencing Calculator** (<http://manisnet.org/gc.html>) to determine "Maximum Uncertainty Distance".

- For **Calculation Type** use
"Error – enter Lat/Long for the actual locality"
- For **Locality Type** use
"Distance along Path". Note: this isn't actually the locality type, but it gives you all of the parameters you need to calculate the correct uncertainty.

LATITUDE AND LONGITUDE COORDINATES

Definition:

The locality consists of a point represented by coordinate information in the form of latitude and longitude. Information may be in the form of

- Degrees, Minutes and Seconds (DMS),
- Degrees and Decimal Minutes (DDM), or
- Decimal Degrees (DD).

Records should also contain a hemisphere (E or W and N or S) or, with Decimal Degrees, minus (-) signs to indicate western and/or southern hemispheres.

Examples:

Example 1: "36° 31' 21.4" N; 114° 09' 50.6" W" (DMS)

Example 2: "36° 31.4566'N; 114° 09.8433'W" (DDM)

Example 3: "36.524276° S; 114.164055° W" (DD)

Example 4: "-36.524276; -114.164055" (DD using minus signs to indicate southern and western hemispheres)

Georeferencing Procedure:

If a location has associated coordinates that are consistent with the rest of the locality description, there is generally little else to do except determine the maximum uncertainty.

Extent:

The extent of a locality should never really be zero. If a GPS was used to determine the coordinates, the accuracy of the GPS at the time (see the section *Using a GPS*, above) should be used as the extent (or see estimates under '*UTM Coordinates*' below. If the coordinates were determined by some other or unknown means, use a reasonable minimum extent for the location, perhaps based on the rest of the locality description. For example, if the coordinates are associated with a point on a trap line, use the distance from the coordinates to the furthest end of the trap line as the extent.

Uncertainty:

Use the *MaNIS Georeferencing Calculator* (<http://manisnet.org/gc.html>) to determine "Maximum Uncertainty Distance".

- For **Calculation Type** use
"Error – enter Lat/Long for the actual locality"
- For **Locality Type** use
"Coordinates Only".

Example 1.

Locality: "35° 22' 24" N, 119° 1' 4" W"

Coordinate System: degrees, minutes, seconds

Latitude: 35° 22' 24" N

Longitude: 119° 1' 4" W

Datum: not recorded; 79 m uncertainty

Coordinate Precision: nearest second; 40 m uncertainty

Coordinate Source: locality description

Distance Units: km, m, mi, yds, or ft

Decimal Latitude: 35.37333

Decimal Longitude: -119.01778

Maximum Uncertainty Distance: 0.119 km, 119 m, 0.074 mi, 130 yds, or 390 ft

Example 2.

Locality: "35.37,-119.02, NAD27, USGS Gosford Quad 1:24000"

Coordinate System: degrees, minutes, seconds

Latitude: 35.27

Longitude: -119.02

Datum: NAD27; no uncertainty

Coordinate Precision: .01 degrees; 1434 m uncertainty

Coordinate Source: USGS map: 1:24,000; 12 m uncertainty

Distance Units: km, m, mi, yds, or ft

Decimal Latitude: 35.37

Decimal Longitude: -119.02

Maximum Uncertainty Distance: 1.446 km, 1446 m, 0.899 mi, 1582 yds, or 4745 ft

UTM COORDINATES

Definition:

The locality consists of a point represented by coordinate information in the form of Universal Transverse Mercator (UTM) or related coordinate system (see Note below). When databasing using UTM or equivalent coordinates, a Zone should ALWAYS be included; otherwise the data are of little value when used outside that zone, and certainly of little use when combined with data from other zones. Zones are often not reported where a region (e.g., Tasmania) falls completely within one UTM zone. Be aware that UTM zones are valid only between 84°N and 80°S.

Note! There are many national and local grids derived from UTM and work in the same way – for example, the Australian Map Grid (AMG).

Examples:

Example 1: “UTM N 4291492; E 456156” (Note: no zone cited).

Example 2: “AMG Zone 56, x: 301545 y: 7011991”

Example 3: “56: 301545.2; 7011991.4”

Georeferencing Procedure:

In Example 1, where no zone is cited, first find the UTM zone by using [UTM Grid Zones of the World](#) (Morton 2006) using any additional information in the locality description, such as country, state/province, county, etc.

Then fill in UTM data at [Geographic/UTM Coordinate Converter](#) (Taylor 2003). Remember that x is easting while y is northing.

Care! Care should be taken when determining UTM coordinates from a map as they are read in the opposite order to Latitude and Longitude, i.e., easting and then northing.

Extent:

See the recommendations under *Latitude and Longitude Coordinates*, above.

Uncertainty:

Calculate in the same way as for *Latitude and Longitude Coordinates*.

If unable to use the Georeferencing Calculator, a general rule of thumb is that the Uncertainty is of the order of

- 30 meters if determined by a GPS after UTC 00:00 2 May 2000, and the datum is recorded;
- 100 meters if determined by a GPS before UTC 00:00 2 May 2000 and the datum is recorded;
- 200 meters plus (depending on the location) if determined by a GPS and the datum is not recorded;
- Variable, depending on map scale if determined from a map (see Table 5, this document).

TOWNSHIP, RANGE, SECTION

Definition:

Township, Range and Section (TRS) or Public Land Survey System (PLSS) is a way of dividing land in the mid- and western USA. Sections are usually 1 mi on each side. Similar subdivisions are used in other countries, and should be calculated in a similar way, once the sizes of the rectangles have been determined. Map sheets are sometimes used and can also be calculates in this way.

A Township Range Section (TRS) description is essentially no different from that of any other named place. It is necessary to understand TRS descriptions and how they describe a place before trying to georeference. See the **References** at the end of this Locality Type, below, for links to further information on Township, Range and Sections and their meaning.

Note! Though TRS applies only to the USA, some countries may have equivalents and the principles elaborated here should be followed.

Examples:

Example 1: "T3S, R42E, SEC.2"

Example 2: "E of Bakersfield, T29S R29E Sec. 34 NE 1/4"

Georeferencing Procedure:

If there is no other usable locality data, or if TRS is the most specific information provided in the locality description, place the point at the center of the TRS or 1/4 section. Otherwise, TRS is best only used as one factor in determining the final coordinates

To find the coordinates for the center of the TRS, use the [TRS-data Website](#) (Gustafson and Wefald 2003) and fill in the appropriate fields. Make sure to pick the correct state. The website will give you the geographic center of your section using the WGS84 datum.

If your locality includes something like "SW .25 Sec 15", then your work is not yet done. To georeference quarters of the section, use the coordinates from the TRS-data website and place them in the Topozone.com website, which shows section boundaries. Put your new point in the appropriate portion of the section and read the new coordinates from the top of the map. Be sure to record the datum used for the coordinates on Topozone®, since these are configurable while looking at the maps

Alternatively, to find the center of a quarter section, first find the center of the Section, then calculate the coordinates of the quarter section by using offsets of 0.25 mi in the appropriate directions from these coordinates. For example, the center of the NW 1/4 of Section 13 would be 0.25 mi N and 0.25 mi W of the center of Section 13.

Note! Not all TRS townships and sections are square. It is best to use a map to find the center of any subdivision of a section.

Extent:

For sections, the extent is half of the hypotenuse of the section, or 0.707 mi (the square root of 2 divided by 2). For quarter sections, the extent is half of that, or 0.354 mi. (Table 6).

Division	Example	Extent (mi)	Extent (m)
Township	T6S R14E	4.243	6828
Section	T6S R14E Sec. 23	0.707	1138
1/4 Section	T6N R14E Sec. 23 NE 1/4	0.354	570
1/4 of 1/4 Section	T6N R14E Sec. 23 NE 1/4 SW 1/4	0.177	285

¼ of ¼ of ¼ Section	T6N R14E Sec. 23 NW ¼ NE ¼ SW ¼	0.089	143
---------------------	---------------------------------	-------	-----

Table 6. Extents of Divisions of Townships in miles and meters. From Wieczorek (2001), Frazier *et al.* (2004).

Uncertainty:

Calculate the same as for '*Feature*'.

If unable to use the Georeferencing Calculator, see Table 6, above. The uncertainty estimate will be the extent plus the uncertainty due to the precision in the coordinates used – as long as the datum is recorded.

References:

Township, Range, Section Information:

<http://www.esg.montana.edu/gl/trs-data.html>

<http://www.outfitters.com/genealogy/land/twprangemap.html>

<http://www.outfitters.com/genealogy/land/land.html>

DUBIOUS

Definition:

At times, locality descriptions are fraught with vagueness. This may be due to any number of reasons, but in particular relates to historic collections in areas that at the time may have had no named features with which to reference.

The most important type of vagueness in a locality description is one in which the locality is in question. Such localities should not be georeferenced.

A cause of vagueness may be incorrect data entry and it is recommended that checking the original catalog books, field notes, specimen labels, etc. be the first step in removing the vaguery of a locality so that it can be georeferenced.

In Examples 1-3, below, the locality descriptions explicitly state that the information contained therein is in question. In Example 4, the location is not well enough bounded to identify a meaningful location. In the latter case, additional information (for example collector and date) may lead to a more specific location from diaries or other published information.

Examples:

Example 1: "possibly Isla Boca Brava"

Example 2: "presumably central Chile"

Example 3: "Bakersfield"

Example 4: "Nova Hollandia"

Georeferencing Procedure:

Do not georeference if the locality explicitly states that the information contained therein is in question.

Document in Remarks the reason for not georeferencing, e.g. "locality too vague to georeference", "locality in question", etc.

Note that subsidiary information may provide other information, which may help in determining a less dubious location.

Extent:

N/A

Uncertainty:

N/A

CANNOT BE LOCATED

Definition:

The cited locality cannot be located. This may be for any number of reasons, including:

- There is no locality information cited (Example 1),
- The locality fields contain other than locality information (Example 2),
- the locality cannot be distinguished from among multiple possible candidates (Examples 3 and 4), or
- the locality cannot be found with available references.

Examples:

Example 1: "locality not recorded"

Example 2: "Bob Jones"

Example 3: ""summit"

Example 4: "San Jose, Mexico"

Georeferencing Procedure:

Do not georeference.

Document in Remarks the reason for not georeferencing, e.g. "locality cannot be found with available references", etc. Do still fill in the Georeferencing Resources field in your database so that the next researcher does not waste time using the same resources to track down the locality.

Extent:

N/A

Uncertainty:

N/A

DEMONSTRABLY INACCURATE

Definition:

The locality contains irreconcilable inconsistencies.

The worst type of locality description to georeference is one that is internally inconsistent. There are numerous possible causes for inconsistencies. Rather than determine coordinates for such localities, annotate the locality with the nature of the inconsistency and refer the locality to the source institution or collector for reconciliation. One common source of inconsistency in locality descriptions comes from trying to match elevation information with the rest of the description. In these cases (see Example 2), bear in mind that elevation data are notoriously inaccurate.

Another common source of inconsistency occurs when the locality description does not match the geopolitical subdivision of which it is supposed to be a part. At times, the locality can still be determined because the geopolitical subdivision is clearly at fault (see Example 3). In this case, georeference the locality and annotate it to describe the problem.

Often there is no way to know if the geopolitical subdivision or something in the locality description itself is at fault. In Example 4, the county may be wrong, the distance may be wrong, or the direction may be wrong. This locality cannot be disambiguated without going back to the originating institution, collection books, or by contacting the collector, etc.

Examples:

Example 1: "Sonoma County side of the Gualala River, Mendocino County"

Example 2: "10 mi W of Bakersfield, 6000 ft" (There is no place anywhere near 10 mi W of Bakersfield at an elevation of 6000 ft)

Example 3: "Delano, Tulare Co." (Delano is in Kern Co.)

Example 4: "5 mi N of Delano, Kern Co." (5 mi N would put the locality in Tulare Co.)

Georeferencing Procedure:

Do not georeference. Record in remarks "locality contains irreconcilable inconsistencies".

Extent:

N/A

Uncertainty:

N/A

CAPTIVE OR CULTIVATED

Definition:

Locality records the collection was made from a captive animal or cultivated plant, etc. The locality cited is often that of a zoo, aquarium, botanical garden, etc.

Examples:

Example 1: "lab born"

Example 2: "bait shop"

Example 3: "Cultivated in Botanic Gardens from seed obtained from Bourke, NSW."

Georeferencing Procedure:

Do not georeference captive animals in standard georeference fields.

Do not georeference cultivated plant records in general georeference fields. If you must supply a georeference (for example for the location of the parent plant that supplied the seed), record it in a separate field, not in the general georeference fields.

Document in remarks "not georeferenced – captive/cultivated", etc.

Extent:

N/A

Uncertainty:

N/A

Glossary

Accuracy — a measure of how well data represent true values.

Cadastral — a register that defines boundaries of public and/or private land.

Cadastral map — a map showing *cadastral* (*q.v.*) boundaries

Coordinates — a sequence of numbers designating the position of a point in n-dimensional space [ISO 19111]. Examples of two-dimensional coordinate systems are Latitude/Longitude and Universal Transverse Mercator (UTM).

Coordinate reference system — a reference system that relates a sequence of numbers or *coordinates* (*q.v.*) to the real world via a *datum* (*q.v.*).

Coordinate system — a system used to denote direct or relative positions by *coordinates* (*q.v.*).

Data Quality — described ‘fitness for use’ (Juran 1964, 1994, Chrisman 1991, Chapman 2005a) of data. As a collector, you may have an intended use for the data you collect but data have the potential to be used in unforeseen ways; therefore, the value of your data is directly related to the fitness of those data for a variety of uses. As data become more accessible, many more uses become apparent (Chapman 2005c).

Datum — a parameter or set of parameters that serve as a reference or basis for the calculation of other parameters [ISO 19111]. A datum defines the position of the origin, the scale, and the orientation of the axes of a coordinate system. A datum may be a geodetic datum, a vertical datum or an engineering datum. In this document, the term *datum* generally refers to a *geodetic datum* (*q.v.*).

Decimal degrees — degrees expressed as a single real number (e.g., -22.343456) rather than as a composite of degrees, minutes, seconds, and direction (e.g., 7° 54' 18.32" E). Note that minus (-) signs are used to indicate southern and western hemispheres.

Decimal latitude — the latitude coordinate (in decimal degrees) at the center of a circle encompassing the whole of a specific locality. Convention holds that decimal latitudes north of the equator are positive numbers less than or equal to 90, while those south are negative numbers greater or equal to -90.
Example: -42.5100 degrees (which is roughly the same as 42° 30' 36" S).

Decimal longitude — the longitude coordinate (in decimal degrees) at the center of a circle encompassing the whole of a specific locality. Decimal longitudes east of the Greenwich Meridian are considered positive and less than or equal to 180, while western longitudes are negative and greater than or equal to -180.
Example: -122.4900 degrees (which is roughly the same as 122° 29' 24" W).

Digital Elevation Model (DEM) — a digital representation of the elevation of locations on the land surface of the earth, usually represented in the form of a rectangular grid.

Eastings and Northings — within a coordinate reference system (e.g., as provided by a GPS or a map grid reference system), *Eastings* are the vertical grid lines running from top to bottom (North to South) which divide a map from East to West and *Northings* are the horizontal lines running from left to right (East to West) dividing the map from North to South. The squares formed by intersecting eastings and northings are called grid squares. On 1:100,000 scale maps each square represents an area of 100 hectares or one kilometer square.

Elevation — the elevation of a geographic location is its height above mean sea level or some other fixed reference point (*cf. vertical datum*). Elevation may be a negative number in those parts of the earth where the land surface is below mean sea level. Elevation may be recorded on maps in the form of contour lines linking points of uniform elevation, or as spot heights at *trig points* (*q.v.*) – usually at the summits of mountains, and rarely at low points. Elevation is used when referring to points on the earth, whereas altitude is used for points above the surface of the earth, such as the altitude of an aircraft, and depth for positions below the surface (of a lake, sea, etc.).

Extent — the geographic range, magnitude, or distance which a location may actually represent. With a town, the extent is the polygon that encompasses the area inside the town’s boundaries. In this document, we usually refer to the linear extent – the distance from the geographic center of the location to the furthest point in the representation of the location.

False Precision — occurs when data are recorded with a greater number of decimal places than implied by the original data. This often occurs following transformations from one unit or coordinate system to another, for example from feet to meters, or from degrees, minutes, and seconds to decimal degrees. In general, precision cannot be conserved across metric transformations; however, in practice it is often recorded as such. For example, a record of 10° 20' stored in a database in decimal degrees is ~10.3°. When exported from some databases, however, it will result in a value of 10.3333333333 with a precision of 10 decimal places rather than 1, leading to a metric uncertainty of around 0.02 mm instead of the real uncertainty of ~15 km. This is not a true precision as it relates to the original data, but a false precision as reported from the database.

Feature — a natural or anthropogenic object or observation that can be represented spatially. The term “feature” may refer to categories of objects or *feature types* (e.g., mountains, roads, or cities) or to specific *feature instances* (e.g., Mount Everest, Interstate 25, or San Francisco), which are also sometimes referred to as “named places.”

Feature Name — a proper name applied to a *feature* (*q.v.*); the name of a named place.

Footprint — a spatial representation of a *feature* (*q.v.*) as an area. The extent and shape of a footprint may comprise the actual boundaries of a feature, the uncertainty around a point representation of a feature, or some combination of an estimate of the boundaries of a feature and the uncertainty associated with those boundaries.

Gazetteer — a geographic dictionary or index of *feature names* (*q.v.*), usually also including an indication of position on the earth’s surface using one of several *geographic coordinate systems* (*q.v.*), but most generally *latitude* (*q.v.*) and *longitude* (*q.v.*).

Geocode — the process of determining the coordinates for a street address. It is also sometimes used as a synonym for **georeferencing** (*q.v.*).

Geodetic datum — a model of the earth used for geodetic calculations. A geodetic datum describes the size, shape, origin, and orientation of a coordinate system for mapping the surface of the earth (NAD27, SAD69, WGS84, etc.). In this document, we use the term to refer to the *horizontal datum* (*q.v.*) and not the *vertical datum* (*q.v.*).

Geodetic datums are often recorded on maps and in gazetteers, and can be specifically set for most GPS devices so the waypoints match the chosen datum. Use "not recorded" when the datum is not known.

Geographic coordinate system — the net or graticule of lines of latitude (parallels) numbered 0° to 90° north and south of the equator, and lines of longitude (meridians) numbered 0° to 180° east and west of the international zero meridian of Greenwich, used to define locations on the Earth's surface (disregarding elevation) with the aid of angular measure (degrees, minutes and seconds of arc)¹⁷.

This is the traditional global coordinate system based on latitude and longitude.

Geographic center — the geographic center of a shape is the mean of the extremes of latitude and longitude of that shape. If the result is not within the shape itself, choose instead the point in the shape nearest to the calculated geographic center.

Georeference — to translate a locality description into a mappable representation of a *feature* (*q.v.*) (verb); or the product of such a translation (noun).

GPS (Global Positioning System) — a satellite-based navigation system that provides 24 hour three-dimensional position, velocity and time information to suitably equipped users (i.e., users with a GPS receiver) anywhere on or near the surface of the Earth. See discussions on accuracy elsewhere in this document.

Heading — the direction from a starting location, given in the form of points of the compass such as E, NW, or N15°W, etc. Usually used in conjunction with *offset* (*q.v.*) to give a distance and direction from a named place. See discussion on true and magnetic north in the **Recording Headings** section of this document.

Horizontal datum — that portion of a *datum* (*q.v.*) which refers to the horizontal positions of mapped features with respect to parallels and meridians or northing and easting grid lines on a map as opposed to the *vertical datum* (*q.v.*).

¹⁷ Glossary of Terminology. <<http://www.ngi.be/NL/glossary/glossang-inf.htm>>

Latitude — describes the angular distance that a location is north or south of the equator, measured along a line of *longitude* (*q.v.*).

Locality — a) the position of a feature in space; b) The verbal representation of this position (i.e., the *locality description*).

Location — a position on the earth's surface or in geographic space definable by *coordinates* (*q.v.*) or some other geographic referencing system, such as a street address, offset, etc.

Longitude — describes the angular distance east or west of a prime meridian (*q.v.*) on the earth's surface along a line of *latitude* (*q.v.*).

Map projection — a method of representing the earth's three-dimensional surface as a flat two-dimensional surface. This normally involves a mathematical model (of which there are many) that transforms the locations of features on the earth's surface to locations on a two-dimensional surface. Such representations distort one or more parameters of the earth's surface such as distance, area, shape, or direction.

Maximum uncertainty estimate — the numerical value for the upper limit of the distance from the coordinates of a locality to the outer extremity of the area (often a circle) within which the whole of the described locality must lie.

Maximum uncertainty units — the units of length in which the maximum uncertainty estimate is recorded (e.g., mi, km, nm, m, ft). The maximum uncertainty distance should be recorded using the same units as the distance measurements in the locality description.

Meridian — the intersection in one hemisphere of the earth's surface with a plane passing through the poles, usually corresponding to a line of *longitude* (*q.v.*).

Named place — used to refer not only to traditional *features* (*q.v.*), but also to places that may not have proper names, such as road junctions, stream confluences, highway mile pegs, and cells in grid systems (e.g., townships).

Northing — See *Eastings and Northings*.

Offset — a displacement from a reference point, named place, or other feature. Used here as the distance from a named place using the location of the named place as the starting point. Usually used in conjunction with *heading* (*q.v.*) to give a distance and direction from a named place.

Precision — with measurements and values, it describes the finest unit of measurement used to express that value (e.g., if a record is reported to the nearest minute, the precision is 1/3600th of a degree; if a decimal degree is reported to two decimal places, the precision is 0.01 of a degree). It is important to always calculate the precision from the original data and units of measurement. See also *false precision* (*q.v.*).

Prime meridian — a meridian from which longitude east and west is reckoned, the most recent standard for which passes through Greenwich, England.

Spatial fit — a measure of how well the geometric representation matches the original spatial representation. See discussion elsewhere in this document.

Trig point — a surveyed reference point, often on high points of elevation (mountain tops, etc.) and usually marked by a small pyramidal structure or a pillar. The exact location is determined by survey triangulation and hence the name trigonometrical point or triangulation point.

Uncertainty — a “measure of the incompleteness of one's knowledge or information about an unknown quantity whose true value could be established if a perfect measuring device were available” (Cullen & Frey 1999). Uncertainty is a property of the observer's understanding of the data. Throughout this document we use *Maximum uncertainty estimate* (*q.v.*) as the way of recording and documenting uncertainty.

UTM (Universal Transverse Mercator) — a standardized coordinate system based on a metric rectangular grid system and a division of the earth into sixty 6-degree longitudinal zones. Zones are numbered consecutively with Zone 1 between 180 and 174 degrees west longitude. UTM only covers from 84° N to 80° S. When citing UTM coordinates, it is essential that the UTM Zone also be recorded.

Vertical datum — that portion of a *datum* (*q.v.*) that refers to the vertical position of mapped features with respect to a base measurement point (such as mean sea level at a location) and from which all elevations are

determined (e.g., AHD – The Australian Height Datum; NAVD88 – North American Vertical Datum). See comments on accuracy under the section on GPS accuracy in this document.

WGS84 (World Geodetic System 1984) — a *coordinate reference system* (q.v.) in common use globally to fit the shape of the entire Earth as accurately as possible using a single ellipsoid. Other ellipsoids (*datums*) are commonly used locally to provide a better fit to the Earth in a local region.

Index to Chapter 5

A

accuracy, iii
 GPS, 9
 index, 37
 of maps, **28**
Alexander Digital Library Gazetteer Server Client, 44
Australian Biodiversity Information Services, 39
Australian Museum, 39

B

best practice
 principles, 1
 accessibility, 1
 accuracy, 1
 effectiveness, 1
 efficiency, 1
 relevance, 2
 reliability, 1
 timeliness, 2
 transparency, 2
Best Practices Guidelines for GPS Survey, 4
BioGeomancer, 1, 18, **19**, 44
 toolkit, 14
BioGeoMancer Classic, **3**, 4, 19, 44
BioGeomancer Consortium, 39

C

cadastral map, iii
cadastre, iii
captive, **76**
caves, **47**
Center for Biodiversity and Conservation (CBC), 46
Centro de Referência em Informação Ambiental, 1, 4, 39
Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, 1, 5
completeness
 index, 36
CONABIO. *See* Comisión Nacional para el Conocimiento y Uso de la Biodiversidad
constraints, **16**
contour line, 52
coordinate precision
 calculating uncertainty from, **27**
coordinate reference system, iii, vi
coordinate system, iii
 geographic, iv
 verbatim, 16
coordinates, iii, v
 geographic, 22
 recording, **8**
 verbatim, 15
CRIA. *See* Centro de Referência em Informação Ambiental
C-squares, 19
cultivated, 76
currency
 index, 37

D

Darwin Core
 Geospatial Element Definitions Extension, 15
data checking, **33**
data cleaning, **33**
data entry, **34**
data entry operators, **20**
data quality, iii, 33
 maintaining, **33**
data validation, **34**
database fields, 14
datum, iii
 engineering, iii
 geodetic, iii, iv, 15
 horizontal, iv, v
 recording, **10**
 unknown, **23**
 vertical, iii, iv, v, vi
datums
 differences between, **23**
decimal degrees, iii, 22
decimal latitude, iii, 15
decimal longitude, iii, 15
DEM. *See* Digital Elevation Model
Denver Botanic Gardens, 3
Denver Museum of Nature and Science, 3
Differential GPS, 9
Digital Elevation Model, iii, 10
direction
 calculating uncertainty from, 26
distance
 calculating uncertainty from, **24**
 precision, 25
DIVA-GIS, 4, 19, 35, 44
documentation, **11**, **37**

E

easting, iii
easting and northing, iii
ecological data, **16**
Ecosystem Associates, 39
elevation, iii
 recording, **10**
Environmental Resources Information Network, 1, **4**, 5, 39
ERIN. *See* Environmental Resources Information Network
error, 33
examples of good and bad localities, 5
extent, iv, **22**
 calculating uncertainty from, **25**
 recording, **11**

F

false precision. *See* precision: false
feature, iv, v, 22, **45**
 between two, **49**

- near, **48**
- subdivisions of, 47
- with extent, 45
- without extent, 46
- feature instances, iv
- feature name, iv
- feature types, iv
- feedback
 - from users, **33**
 - to collectors, **33**
- fitness for use, iii
- footprint, iv
- Fuzzyg - Fuzzy Gazetteer, 44

G

- gazetteer, iv
- GBIF. *See* Global Biodiversity Information Facility
- GeoCalc, 44
- geocode, iv
- geodetic datum. *See* datum: geodetic
- geographic center, v
- geographic coordinate system, iv
- geographic coordinates, 22
- Geographic Information System, 22
- geographic regions, 19
- GeoLoc - CRIA, 44
- GEOLocate, 1, **4**, 5, 19, 44
- GEOnet Names Server, 44
- georeference, v
 - determined by, 16
 - determined date, 16
 - protocol, 16
 - remarks, 16
 - sources, 16
 - validation, 16
 - verification status, 16
- georeferencing
 - batch, **18**
 - beginning, **13**
 - fields, 15
 - legacy data, **21**
 - methodology, **18**
- Geospatial Element Definitions Extension to Darwin Core, 15
- Global Biodiversity Information Facility, 12, 39
 - Portal, 19, 35
- Global Gazetteer, 44
- Global Positioning System, v, 7, 8
 - accuracy, 9
 - Differential, 9
 - Local Area Augmentation System, 9
 - Real Time Differential, 9
 - Static, 9
 - using, **8**, 12
 - Wide Area Augmentation System, 9
- Globally Unique Identifier, 12
- Gordon and Betty Moore Foundation, 39
- GPS. *See* Global Positioning System
- Greenwich Meridian, iii
- GUID. *See* Globally Unique Identifier
- Guide for Recording Localities in the Field, 5
- guidelines, **17**

H

- heading, v
 - recording, **10**
- Herbarium Information Standards and Protocols for Interchange of Data, 15
- HerpNet, 30

I

- index
 - accuracy, 37
 - completeness, 36
 - currency, 37
 - validation, 37
- Index of Spatial Uncertainty, **36**
- INRAM. *See* Institute of Resource Analysis and Management
- Institute of Resource Analysis and Management, 1, **3**, 5
- International Rice Research Institute, 39

J

- junction
 - road, 46

L

- latitude, iv, v, **22**
 - decimal, iii, 15
- legacy data
 - georeferencing, **21**
- Local Area Augmentation System, 9
- locality
 - database fields, 14
 - recording, **7**
- locality description
 - classifying, **21**
- Locality Type
 - Between Two Features, **49**
 - Between Two Paths, **53**
 - Cannot be Located, **74**
 - Captive or Cultivated, **76**
 - Demonstrably Inaccurate, **75**
 - Dubious, **73**
 - Feature, 22, **45**
 - Latitude and Longitude Coordinates, **68**
 - Named Place. *See* Locality Type: Feature
 - Near a Feature, **48**
 - Offset Along a Path, **62**
 - Offset at a Heading, **60**
 - Offset Direction, **56**
 - Offset Distance, **54**
 - Offset from Two Distinct Paths, **67**
 - Offset in Orthogonal Directions, **64**
 - Path, 8, **51**
 - Street Address, **50**
 - Township, Range, Section, **71**
 - UTM Coordinates, **70**
- location, v
- longitude, iv, v, **22**
 - decimal, iii, 15

M

- magnetic declination
 - calculating, **11**
- Magnetic North, **10**
- making corrections, **35**
- Mammal Networked Information System, **1, 3, 4, 5, 13, 18, 30**
- manager
 - responsibilities of, **35**
- MaNIS. *See* Mammal Networked Information System
- MaNIS Georeferencing Calculator, **19, 27, 30, 44**
 - Manual, **30**
- MaNIS/HerpNet/ORNIS Georeferencing Guidelines, **5, 30**
- map
 - accuracy of, **28**
 - calculating uncertainty from, **28**
- map projection, **v**
- map squares, **19**
- MaPSTeDI. *See* Mountains and Plains Spatio-Temporal Database Informatics Initiative
- MapSTeDI Georeferencing Guidelines, **34**
- MaPSTeDI Georeferencing Protocols, **3, 5**
- MaPSTeDI Guide to Georeferencing, **3, 5**
- maximum uncertainty
 - estimate, **v, vi, 15**
 - unit, **v, 15**
- meridian, **v**
 - Greenwich, **iii**
 - prime**, **vi**
- Mountains and Plains Spatio-Temporal Database Informatics Initiative, **1, 3, 5, 13, 14, 18, 34**
- Museum of Vertebrate Zoology, **5**
- MVZ Guide for Recording Localities in the Field, **5**

N

- named place, **v, 45**
- National Geophysical Data Center, **11**
- NatureServe, **1, 39**
- NGDC Magnetic Declination Calculator, **11, 44**
- North
 - Magnetic, **10**
 - True, **10**
- northing, **iii, v**

O

- offset, **v, 22**
 - along a path, **62**
 - at a heading, **60**
 - from two distinct paths, **67**
 - in orthogonal directions, **64**
- offset direction, **56**
- offset distance, **54**
- OGC Recommendations, **5**
- ORNIS. *See* ORNithological Information System
- ORNithological Information System, **30**

P

- path, **8, 51**

- between two, **53**
 - subdivision of, **52**
- PDA. *See* Personal Digital Assistant
- performance criteria, **36**
- Perpendicular Distance Calculator, **46**
- Personal Digital Assistant, **12**
- precision, **vi**
 - distance, **25**
 - false, **iv, vi, 27**
- prime meridian**, **vi**
- properties, **47**
- Public Land Survey System, **71**

R

- recording
 - coordinates, **8**
 - data for small labels, **12**
 - datum, **10**
 - elevation, **10**
 - extent, **11**
 - headings, **10**
 - localities, **7**
 - year of collection, **11**
- rivers, **51**
- roads, **51**

S

- Selective Availability, **8, 9**
- small labels
 - recording data, **12**
- spatial fit, **vi, 16, 31**
- standards, **17**
- supervisor
 - responsibilities of, **35**

T

- Taxonomic Databases Working Group, **12, 39**
- TDWG. *See* Taxonomic Databases Working Group
- Township, Range and Section, **71**
- training, **36**
- trig point, **iv, vi**
- True North, **10**
- Truth in Labelling, **35**
- Tulane University, **39**
 - Museum of Natural History Fish Collection, **4**

U

- uncertainty, **v, vi**
 - calculating, **23**
 - calculating combined, **30**
 - calculating due to unknown datum, **23**
 - calculating from a map, **28**
 - calculating from coordinate precision, **27**
 - calculating from direction, **26**
 - calculating from distance, **24**
 - calculating from extents, **25**
- Universal Transverse Mercator, **vi, 70**
- University of California
 - Berkeley, **39**

Merced, 39
University of Colorado Museum, 3
University of Colorado, Boulder, 39
University of Illinois Urbana-Champaign, 39
University of New Mexico, 39
University of Tulane, 5
unknown datum
 uncertainty due to, **23**
user interfaces, **17**
UTM. *See* Universal Transverse Mercator

V

validation

index, 37

W

WGS84. *See* World Geodetic System 1984
Wide Area Augmentation System, 9
World Geodetic System 1984, vi

Y

Yale University, 39
year of collection
 recording, **11**

Chapter 6

Generalising Sensitive Data

Introduction.....	1
Principles.....	2
Determining sensitivity	3
CRITERIA FOR DETERMINING SENSITIVITY	3
CATEGORIES OF SENSITIVITY	7
Listing sensitive taxa.....	10
Generalising textual information.....	11
Generalising spatial information	12
DOCUMENTATION.....	13
Documentation and metadata.....	13
DOCUMENTING SENSITIVITY	14
Authentication and authorisation	16
References	17
Appendix: Scenarios using criteria 1 and 2 as triggers	18
Glossary	20

This Chapter is equivalent to:

Chapman, A. D. and O. Grafton. 2008. *Guide to Best Practices for Generalising Primary Species-Occurrence Data*, version 1.0. Copenhagen: Global Biodiversity Information Facility. 27 pp. ISBN: 87-92020-06-2 (available as a standalone PDF from <http://www.gbif.org>)

Introduction

The unprotected distribution of Sensitive Primary Species Occurrence Data (for example the exact localities of rare, endangered or commercially valuable taxa) has been a concern of the [GBIF Secretariat](#) since its beginning. In early 2006, GBIF initiated a process to address this issue, especially in relation to data to be shared through the GBIF network and made visible through the [GBIF Data Portal](#).

A review of current approaches for obscuring or generalising such data was initiated in February 2006 and an on-line survey conducted through Survey Monkey¹. A separate report on the results was made available via the GBIF Web site² in early June 2006 (Chapman 2006). An experts' workshop was then held in early March 2007 that focussed on the various technical issues involved (Chapman 2007a).

A final report on Dealing with Sensitive Primary Species Occurrence Data was developed following these processes and discussions, and was presented to GBIF in April 2007 (Chapman 2007b). It is available via the GBIF Web site. This report made a number of recommendations, and many of these are included in this document.

The final step in this process has been to develop a Guide to Best Practices. This document should be seen as an overriding guideline for institutions, data providers and GBIF Nodes to use to develop their own in-house guidelines. Organisations and institutions should produce their own internal document that incorporates the practices outlined in this document and related documents such as the Guide to Best Practices in Georeferencing (Chapman and Wieczorek 2006) and incorporate them into their own working environment.

It is also important to understand the possible impact that approaches for restricting sensitive data may have on biodiversity science and, while restricting the availability or resolution of certain data, not overly restricting the uses to which the data may be put. For that reason, a set of principles are elucidated below. Key among these is the need to make biodiversity information freely available wherever possible, in the interests of science, the environment and the biodiversity itself.

Two issues that this document has not covered, because they will need further discussion and agreement before robust recommendations can be made, are the issues of the privacy of living individuals and the development of Data Sharing and Data License Agreements. Both of these issues have legal implications and vary considerably from jurisdiction to jurisdiction. Recommendations were made in the Report on Dealing with Sensitive Species Occurrence Data (Chapman 2007) for GBIF to further explore these issues.

“The term best practice generally refers to the best possible way of doing something; it is commonly used in the fields of business management, software engineering, and medicine, and increasingly in government. [...] The [qualified] term, ‘best current practice’, often represents the meaning in a more accurate way, showing the possibility for future developments of ‘better practice’.”

[\(Wikipedia: Best Practice\)](#).

¹ Survey Monkey <http://www.surveymonkey.com>

² http://www.gbif.org/prog/digit/sensitive_data/Summary_of_Responses_-_03.pdf

Principles

Biodiversity information should be made freely available to be shared globally to enable their use for not-for-profit decision-making, education, research and other public benefit purposes. Making the full detail of biodiversity information available should reduce the risk of damage to the environment and help safeguard a sustainable future. Where release will have the opposite effect, access to the full detail may need to be controlled.

Below are a set of high level principles related to the sharing of data generally, and the sharing of sensitive data in particular.

- 1. Wherever possible, environmental information should be freely available to all. Generally this benefits the environment by increasing awareness, enabling better decision-making and reducing risk of damage.*
- 2. In a small number of cases, public access to information can result in environmental harm. It should be recognised that in such cases, availability of information may need to be controlled; although the presumption remains in favour of release and any restrictions should be interpreted rigorously.*
- 3. All data regarded as being sensitive should include a date for review of their sensitivity status, along with documented reasons for the sensitivity status. The date for review may be short or long depending on the nature of the sensitivity. Whenever a data provider receives an application for enhanced access to restricted data they should avoid assuming continued sensitivity and use it as an opportunity to revisit the determination.*
- 4. If the data are to be restricted for distribution, then this should only be done to a copy of the data at the time of their distribution. Data should never be altered, falsified or deleted from the stored record.*
- 5. Documentation is essential for many reasons, and where data have been restricted or generalised it is important that that information is recorded as metadata that remains with the record.*
- 6. Where data are restricted or generalized for distribution (such as the name of a collector, textual locality information, etc.) this should be documented by replacing with appropriate wording – the field should not be left blank or null.*
- 7. There are extremely strong reasons not to restrict data on related collections (e.g. collector's numbers in sequence, collector's name, etc.) because of the restrictions this places on data quality/ data validation procedures and the limits it places on the effectiveness of filtered Push Technologies.*
- 8. Users of sensitive data should respect any and all restrictions of access that the data provider has placed on the data. If granted enhanced access to restricted information users must not compromise or otherwise infringe the confidentiality of such information.*
- 9. Data providers should respect the needs of data users to have access to data and documentation in order to determine the 'fitness for use' of the data, and to ensure that analyses are robust and not misleading.*

Determining Sensitivity

As a first step, information holders need to identify any data which are regarded as 'sensitive'. Sensitive information is any which if released to the public, would result in an 'adverse effect' on the taxon or attribute in question or to a living individual. A number of factors need to be taken into account when determining sensitivity, including type and level of threat, vulnerability of the taxon or attribute, type of information, and whether it is already publicly available. Determining these factors leads us to a criteria-based approach.

Two examples of sensitivity criteria that provide a starting point for the development of criteria are those developed by the National Biodiversity Network (NBN) in the UK (National Biodiversity Network 2002, 2004), and the Department of Environment and Conservation in New South Wales, Australia (Department of Environment and Conservation 2007).

Below are a series of criteria for determining the sensitivity of taxa and data along with recommended metadata statements for documenting the reasons for the determination. The first two are for use by biodiversity data holders and those creating trigger lists of potentially sensitive taxa and refer largely to the taxa themselves. The last two are for use by biodiversity data holders and deal with an assessment of the data they hold and are considering making available – they are not suitable for the creation of trigger lists.

The criteria are used to determine:

1. Risk of Harm	An assessment of whether the taxon is subject to harmful human activity.
2. Impact of Harm	An assessment of the sensitivity of the taxon to the harmful human activity.
3. Sensitivity of Data	An assessment on whether the release of data will increase harm.
4. Decision on release & Category of sensitivity	A balanced decision regarding the release of the data and a determination of the category of sensitivity, and thus the level of generalisation, of the data for release.

A set of scenarios using Criteria 1 and 2 below for determining triggers for sensitivity of taxa is attached as an Appendix to this chapter.

Criteria for Determining Sensitivity

The first step in the process of determining sensitivity is to make an assessment on whether or not the taxon is subject to a harmful human activity and if the availability of related biodiversity data will increase the likelihood of the harmful activity occurring. If it is not then there would appear no reason to list it as a potential environmentally sensitive taxon. It is recommended that you use the documented wording supplied but with additional supporting rational documenting the specifics of the threat, for example:

“The taxon is at risk from harmful human activity –it is subject to attack by Phytophthora which is transported by human operated vehicles.”

1. RISK OF HARM

Assess whether the taxon is subject to a harmful human activity.	
1.1. Is the taxon subject to a harmful human activity?	Yes: Document using statement <i>1a</i> with supporting rationale. Go to 1.2
	No: Document using statement <i>1b</i> [Taxon is not sensitive] Go to 3
1.2. Is there established evidence of current or recent occurrences of the harmful human activity?	Yes: Document using statement <i>1c</i> with supporting rationale. Go to 1.3
	No: Document using statement <i>1d</i> with supporting rationale. Go to 1.3
1.3. Will availability of related biodiversity data increase the likelihood of the harmful human activity taking place?	Yes: Document using statement <i>1e</i> with supporting rationale. Go to 2
	No: Document using statement <i>1f</i> with supporting rationale. Go to 2

1a – The taxon is at risk from a harmful human activity.
1b – There is no significant risk of a harmful human activity.
1c – There is established evidence of actual or recent harm to the taxon
1d – There is currently no established evidence of actual harm to the taxon.
1e – Availability of biodiversity data will increase the likelihood of the harmful human activity taking place.
1f – Availability of biodiversity data will not increase the likelihood of the harmful human activity taking place.

The next step is to determine if the taxon is sensitive to that human harm or whether it is suitably robust so as not to be adversely affected.

2. IMPACT OF HARM	
Assess the sensitivity of the taxa to the harmful human activity.	
2.1. Does the taxon have characteristics that make it significantly vulnerable to the harmful human activity?	Yes: Document using statement <i>2a</i> with supporting rationale. Go to 2.2
	No: Document using statement <i>2b</i> and supporting rationale. Go to 2.2
2.2. Is the taxon vulnerable to harmful human activity over its total range, or are there areas (such as in conservation zones, or other parts of the world) where the taxon is not at the same level of risk?	Yes: Document using statement <i>2c</i> with supporting rationale. Go to 3
	No: Document using statement <i>2d</i> with supporting rationale. Go to 3

2a – The taxon has characteristics that make them significantly vulnerable to the harmful human activity.
2b – The taxon is not significantly vulnerable to the harmful human activity.
2c – The taxon is vulnerable to harmful human activity over its total range.
2d – The taxon is not vulnerable to harmful human activity over its total range and/or there are areas where the taxon occurs but is not at significant risk.

Once it has been decided that the taxon is subject to a significant risk and impact from harm or not, then a decision needs to be taken on whether the release of specific data on that taxon – or other related data – will increase the risk and impact of harm.

3. SENSITIVITY OF DATA	
Assess whether the release of data will increase harm.	
3.1. Is the content and detail of the biodiversity data such that their release would enable someone to carry out a harmful activity upon the taxon or attribute?	Yes: <i>Document using statement 3a with supporting rationale.</i> Go to 3.2
	No: [Data are not sensitive] <i>Document using statement 3b with supporting rationale</i> Go to 4
3.2. Is information already in the public domain, or already known to those individuals or groups likely to undertake the harmful activity?	Yes: <i>Document using statement 3d with supporting rationale.</i> Go to 3.3
	No: <i>Document using statement 3c with supporting rationale.</i> Go to 3.3
3.3. Would disclosure damage a partnership or relationship (especially where the maintenance of which is essential to helping achieve a specific conservation objective)?	Yes: <i>Document using statement 3e with supporting rationale.</i> Go to 3.4
	No: <i>Document using statement 3f with supporting rationale.</i> Go to 3.4
3.4. Would disclosure allow the locations of sensitive features to be derived through combination with other publicly available information sources?	Yes: <i>Document using statement 3g with supporting rationale.</i> Go to 4
	No: <i>Document using statement 3h with supporting rationale.</i> Go to 4

3a – The content and detail of the data is such that their release would enable someone to carry out a harmful activity upon the taxon or attribute.
3b – The content and detail of the data if released would not enable someone to carry out a harmful activity upon the taxon or attribute.
3c – The information is <u>not</u> in the public domain, and is not already known to individuals or groups likely to undertake harmful activities.
3d – The information is already in the public domain, or is already known to the individuals or groups likely to undertake harmful activities.
3e – Disclosure of the data is <u>likely</u> to damage a partnership or relationship the maintenance of which is essential to helping achieve a specific conservation objective.
3f – Disclosure of the data <u>will not</u> damage any partnership or relationship essential to conservation.
3g – Disclosure <u>would</u> allow the locations of sensitive features to be derived through combination with other publicly available information sources
3h – Disclosure <u>will not</u> allow the locations of sensitive features to be derived through combination with other publicly available information sources

The final step is to make an overall assessment based on the three criteria above and to document the overall decision using the combined information documented in making each

of the earlier decisions. Once it has been determined that the data should or should not be released, then it is important that a decision is made on the Category of Sensitivity, and the level of generalisation for the release of the data.

4. DECISION ON RELEASE & CATEGORY OF SENSITIVITY	
Make a balanced decision regarding the release of data and determining the category and level of generalisation	
4.1. On balance, considering criteria 1 to 3 above and any important wider context, will withholding the information increase the risk of environmental harm or harm to a living person?	Yes: <i>Document using statement 4a.</i> Go to 4.2
	No: <i>Document using statement 4b.</i> Go to 4.5
4.2. Is the taxon distinctive and of high biological significance, under high threat from exploitation/ disease or other identifiable threat where even <u>general</u> locality information may threaten the taxon? Or could the release of any part of the record cause <u>irreparable harm</u> to the environment or to an individual?	Yes: <i>Document using statement 4c, collate all supporting rationale and document the decision to withhold the data.</i> Go to Category 1
	No: Go to 4.3
4.3. Is the taxon such that the provision of precise locations at finer than 0.1 degrees (~10 km) would subject the taxon to threats such as disturbance and exploitation? Or does the record include highly sensitive information, the release of which could cause <u>extreme harm</u> to an individual or the environment?	Yes: <i>Document using statement 4d, collate all supporting rationale and document the decision to release the data.</i> Go to Category 2
	No: Go to 4.4
4.4. Is the taxon such that the provision of precise locations at finer than 0.01 degrees (~1 km) would subject the species to threats such as collection or deliberate damage? Or does the record include sensitive information, the release of which could cause <u>harm</u> to an individual or the environment?	Yes: <i>Document using statement 4e, collate all supporting rationale and document the decision to release the data.</i> Go to Category 3
	No: Go to 4.5
4.5. Is the taxon subject to low to medium threat if precise locations (i.e. locations with a precision greater than 0.001 degrees or 100m) become publicly available and where there is some risk of collection or deliberate damage?	Yes: <i>Document using statement 4f, collate all supporting rationale and document the decision to release the data.</i> Go to Category 4
	No: <i>Document using statement 4g, collate all supporting rationale and document the decision to release the data.</i> Data should be publicly released

4a – On balance, release of the information will, or is likely to, increase the risk of environmental harm or harm to a living person.
4b – On balance, release of the data will not increase the risk of environmental harm or harm to a living person.
4c – The species is a distinctive species of high biological significance, is under high threat from exploitation/ disease or other identifiable threat and even general locality information may threaten the taxon, or the release of the information could cause irreparable harm to the environment, an individual, or some other feature. [Category 1]
4d – The species is classed as highly sensitive, and the provision of precise locations would subject the species to threats such as disturbance and exploitation, and/or the record includes highly sensitive information, the release of which could cause extreme harm to the environment or an individual. [Category 2]
4e – The species is classed as of medium to high sensitivity, and the provision of precise locations could subject the species to threats such as collection or deliberate damage, and/or the record includes sensitive information, the release of which could cause harm to the environment or to an individual. [Category 3]
4f – The species is classed as of low to medium sensitivity, and the provision of precise locations could subject the species to threats such as disturbance and exploitation. Detailed data may be made available to individuals under license. [Category 4]
4g – The species is classed as of low sensitivity, and the distribution of precise locations is unlikely to subject the species to significant threat, and/or the record includes information of low sensitivity, the release of which is unlikely to cause harm to the environment or to any individual. The data should be released to the public ‘as-held’ [Not Environmentally Sensitive]

In the on-line survey, a number of respondents identified data awaiting publication, data subject to ongoing research, and incomplete or unchecked data as data that they would class as sensitive, and thus subject to restrictions on release. These are data whose sensitivity has a short time frame and it is important that a time for release or review be clearly documented. They would most likely fall under criterion 3.3 above and would be documented accordingly with the supporting rationale being “*awaiting publication*”, etc.

NB. All data regarded as being sensitive, should include a date for review of their sensitivity status, along with documented reasons for the sensitivity status. The date for review may be short or long depending on the nature of the sensitivity.

The Categories of Sensitivity (below) are largely based on those from the New South Wales Department of Environment and Conservation (2007).

1. Categories of Sensitivity

Criterion	Reasoning
Category 1 –Species or records for which no records will be provided at all, or which are only	The reason for non-disclosure is that: 1. a <i>distinctive</i> species of high biological significance is under high threat from exploitation/ disease or other identifiable

Criterion	Reasoning
<p>released as present within a large region such as a county, watershed, etc.</p>	<p>threat where even general locality information may threaten the taxon.</p> <p>2. the information in the record is of such a nature that its release could cause irreparable harm to the environment, to an individual or to some other feature.</p> <p>Data may only be supplied under strict License conditions or as presence in a large region such as a watershed, county, or biogeographic region.</p>
<p>Category 2 – Species or records for which coordinates will be publicly available ‘denatured’ (to 0.1 degrees) and/or other information in the record is generalized. Finer scale data (Category 3 or 4 or detailed data) may be supplied to individuals under License.</p>	<p>The reasons for restriction are that:</p> <ol style="list-style-type: none"> 1. The species is classed as <i>highly sensitive</i>, and the provision of precise locations <i>would</i> subject the species to threats such as disturbance and exploitation. 2. The record includes <i>highly</i> sensitive information, the release of which could cause <i>extreme</i> harm to an individual or to the environment. <p>Data are supplied to the public</p> <ol style="list-style-type: none"> 1. with the georeference denatured to 0.1 degrees (~10 km) and/or 2. with sensitive fields generalized or removed and replaced with suitable replacement wording. <p>Data may be supplied at finer scales on request under the conditions of a written data agreement, usually a Data Licence Agreement. When data are provided to clients, they will be advised which species or fields are sensitive and may have their coordinates denatured to that available under Categories 3 or 4.</p> <p>NB. In the case where the sensitivity is triggered by fields other than the georeference, it may be more appropriate to class the record as Category 3 or 4.</p>

Criterion	Reasoning
<p>Category 3 – Species or records for which coordinates will be publicly available ‘denatured’ (to 0.01 degrees) and/or other information in the record is generalized. Finer scale data (Category 3 or 4 or detailed data) may be supplied to individuals under License.</p>	<p>The reasons for restriction are that:</p> <ol style="list-style-type: none"> 1. The species is classed as of <i>medium to high sensitivity</i>, and the provision of precise locations <i>could</i> subject the species to threats such as disturbance and exploitation. 2. The record includes <i>sensitive</i> information, the release of which could cause harm to an individual or to the environment. <p>Data are supplied to the public</p> <ol style="list-style-type: none"> 1. with the georeference denatured to 0.01 degrees (~ 1 km) and/or 2. with sensitive fields generalized or removed and replaced with suitable replacement wording. <p>Data may be supplied at finer scales on request under the conditions of a written data agreement, usually a Data Licence Agreement. When data are provided to clients, they will be advised which species or fields are sensitive and may have their coordinates denatured to that available under Category 4.</p> <p>NB. In the case where the sensitivity is triggered by fields other than the georeference, it may be more appropriate to class the record as Category 4.</p>
<p>Category 4 – Species or records for which coordinates will be publicly available ‘denatured’ (to 0.001 degrees) and/or other information in the record is generalized. Detailed ‘as-held’ data may be supplied to individuals under License.</p>	<p>The reasons for restriction are that:</p> <ol style="list-style-type: none"> 1. The species is classed as of <i>low to medium sensitivity</i>, and the provision of precise locations could lead to risk of collection or deliberate damage. 2. The record includes <i>sensitive</i> information, the release of which could cause harm to an individual or to the environment. <p>Detailed data may be supplied under the conditions of a written data agreement, usually a Data Licence Agreement. When data are provided to clients, they will be advised which species or fields are sensitive.</p>

Listing Sensitive Taxa

Data are already distributed around the globe through duplicate specimens, etc. and although data may be restricted from some institutions, others holding duplicates may be releasing the same information. This may be through ignorance of what may be regarded as sensitive in the home ranges of the taxon concerned as no universal list of what is regarded as 'sensitive' is currently available. Difficulties are compounded by the fact that a taxon may be sensitive in one area, but not in another (and indeed may even be a weed or pest species in the second location).

For these reasons it has been recommended that a trigger list of potential environmentally sensitive taxa should be created and linked through GBIF's Electronic Catalogue (ECat)³. This would have the advantages of alerting data providers in other jurisdictions that a species is potentially sensitive, and via ECat would provide links to synonyms. It is important to note that the list should be regarded as a trigger to flag the need for a decision on the actual sensitivity of sharing information using the criteria in the previous chapter, and not for generating blanket restrictions. Not all endangered species are threatened through knowledge of their locations and so should not be regarded as sensitive *per se* and thus the list of potential environmentally sensitive taxa should be much smaller than any existing list of rare and threatened species.

The list should be created using Criteria 1 and 2 (refer to the previous Chapter and scenarios in Annex 1), and should include additional information such as:

- Name of Taxon
- Criteria and supporting rationale for inclusion
- Name of person or organisation responsible for the taxon being included
- Geographic coverage of sensitivity (especially if only sensitive over part of its range or within one jurisdiction)
- Recommended Sensitivity Category
- Date for Review

Jurisdictions may also wish to maintain a similar list for their own purposes, and it is recommended that if they do so, they include the above information in all cases. The advantages of making the information more broadly available is that it will alert other data custodians that your jurisdiction regards the taxon as potentially sensitive, and alert users that they should take the sensitivity into account when publishing the results of their analyses, etc.

NB. *Any list of potential environmentally sensitive taxa should be regarded as a trigger only, and any restrictions on availability of actual data should be made on a case by case basis taking into account the listed criteria.*

³ GBIF Electronic Catalogue <http://www.gbif.org/prog/ecat>

Generalising Textual Information

In some cases, information in text fields might be regarded as sensitive under certain circumstances. This may include such information as:

- Names of living persons
- Locality information
- The date of collection
- The collector's number
- Habitat
- Landholder information
- Taxonomic names

Some of these may need to be restricted to stop co-relational analyses leading to deductions on the localities of records that are restricted or generalized – for example the collector's name, date, and collector's numbers in sequence. In other cases, it may be necessary to hide the name of a taxon in a list of collections in a biodiversity hot-spot or sensitive locality.

Such restrictions should not restrict the provision of the record as a whole. The data that needs to be hidden may be removed and replaced with suitable wording (see below), or generalized – for example, just giving the name of a higher level taxonomic rank where the species is to be restricted.

NB. *Whenever data in a textual field are restricted or generalized for distribution (such as the name of a collector, textual locality information, etc.) it should be documented by replacing it with appropriate wording – the field should **not** be left blank or null.*

Examples of replacement wording include:

- *“name suppressed for reasons of privacy”;*
- *“This specimen represents an endangered or threatened species. The specific locality has been removed from the on-line record to protect this species from over-collection. These data may be supplied to researchers on request”;*
- *“This specimen represents an endangered or threatened species. The specific locality has been generalized to presence within a grid of 0.1 degree resolution. Detailed data may be supplied to researchers on request”.*

NB. *Where there is need to restrict a taxonomic name (for example, of sensitive taxa as part of a survey), it may be possible to replace it with a higher taxon name (genus/family, etc.), or to just report that there are ‘x’ sensitive taxa present without providing names.*

Occasionally, data providers may be tempted to restrict information in records related to a sensitive record (in addition to the sensitive record itself), such as the collector's name and numbers in a sequence of records collected at the same location and time as a sensitive record in order to reduce the possibility of the sensitive record being found through co-relational analysis. However, if the collector's name and number is removed from just the sensitive

record and not the others, it is unlikely that these could be deduced unless the seeker of the information already has inside knowledge. For this reason, and others (see box below), it is recommended that the data on related records not be restricted.

NB. *There are extremely strong reasons not to restrict data on related collections (collector's numbers in sequence, collector's name, habitat, etc.) because of the restrictions this places on data quality/ data validation procedures and the limits it places on the effectiveness of filtered Push Technologies. Information in records related to a sensitive record (but not in the sensitive record itself) should not be restricted unless absolutely necessary.*

Generalising Spatial Information

One of the most common requirements for generalising biodiversity information is to generalise the spatial locality or georeference. Traditionally this has been done in many ways, and there has been little consistency in methodologies, and very little documentation as to what has been done in each case. This has considerably reduced the value of the data for analysis, and often users are unaware that the data has even been modified.

Good practice dictates that whatever you do to generalise the data that you document it so that users of the data know what reliance they can place in them.

Following considerable discussion among data providers and data users, it has been decided to recommend that data providers who are generalising their data do so using a standard methodology (see below), and to document this accordingly. As most biodiversity data are currently made available using decimal degrees, the recommended method means that protocols (such as Darwin Core) do not need modification, other than to allow for suitable metadata documentation.

The method recommended below allows for several levels of generalisation that conform to Categories 1-4 described in the earlier Chapter on *Determining Sensitivity*.

The recommended method for generalisation is:

Category	Sensitivity	Georeference
Category 1	Extreme	Georeference not released or data may be released by watershed/ bioregion/ county, etc. with no georeference coordinates.
Category 2	High	Georeference rounded to 0.1 degree
Category 3	Medium	Georeference rounded to 0.01 degree
Category 4	Low	Georeference rounded to 0.001 degree
Not sensitive	Not sensitive	Georeference unrestricted.

Documentation

It is important to document the method and level of generalisation so that users are aware of what has been done to the data, and what reliability they may be able to place in the data. Currently, neither Darwin Core nor the ABCD protocols provide fields for the recommended metadata. It has been recommended, however, that these protocols be modified to accept such metadata (see Chapter on *Documentation and Metadata*), but in the meantime, it is recommended that the information be recorded in Comments fields.

As far as the generalisation of georeferencing data is concerned it is important to record that the data has been generalized using a 'decimal geographic grid', and record both:

- Precision of the data provided (e.g. 0.1 degree; 0.001 degree, etc.)
- Precision of the data stored or held (e.g. 0.0001 degree, 0.1 minute, 1 second, etc.)

The recommendations for metadata for inclusion in the [Geospatial Element Definitions Extension](#) to Darwin Core (TDWG 2005) are set out in the next Chapter on *Documentation and Metadata*. Once they (or similar) have been adopted, then it is recommended that the appropriate fields be recorded and distributed with the data.

NB. If generalizing to a large region such as a watershed, biogeographic region or a county, etc., then do not supply a georeference.

Documentation and Metadata

It is important that data be accurately documented so that users and others know exactly what the data represent, and the reliance that can be placed in them. For example, a user needs the information to determine if the data are suitable for the analysis they are about to run. Many data providers reported in the survey that one reason that they were reluctant to release some of their data was a fear that the data would be mis-used. If the data aren't adequately documented, then the likelihood of inadvertent mis-use is greatly increased as the user may use the data in an analysis mistakenly thinking they are getting accurate point records, when in reality, the data had been generalized to a 10 km grid square, and could be anywhere in a 100 square kilometre area. If running a climate modelling algorithm, for example, then this sort of error could result in a quite misleading result. For this reason alone, it is important to data providers, data users, and end users (such as environmental managers, policy makers, etc.) that the data are accurately described.

In particular, there should be a clear documentation of the 'Access Constraints' which could include, for example, an indication of which parts of the data are sensitive (if any), reasons for sensitivity and conditions under which release is possible.

Documenting Sensitivity

“Metadata fulfils an essential function regarding communication to third parties, of access constraints and use conditions that the data generators intend to give to their data. It can be considered as an ‘aid’ in protecting data and information, since it will allow system users to visualize the conditions established by the data generator for access and use of the information. Additionally, in case the data are not accessible, the metadata allows knowledge of the conditions of access through other media (digital or not) as well as a summary of the content”. (Llinás, 2005).

Metadata has generally been used to refer to documentation of a whole dataset. Documentation at the record level has usually been referred to just in comments. I prefer, however, to term this ‘*record-level metadata*’ (see glossary) and to formalise the process. In the previous chapter a recommendation was made that where data were generalized for distribution, to document the level of generalisation - for example, that the data had been generalized using a decimal geographic grid, and to record both the precision of the data provided and the precision of the data stored or held. Also, in the chapter on *Determining Sensitivity*, a series of documentation processes were recommended. Some of these may be more appropriate for documenting the reasons for regarding a taxon as a potential environmentally sensitivity taxon (Criteria 1 and 2), while the others (Criteria 3 and 4) are appropriate to the data themselves and belong as part of the broader record-level metadata. To fully document the reasons for restricting data, however, it may be necessary to inherit the documentation from Criteria 1 and 2 to the record level – for example, the reason that data are restricted may include that the taxon is subject to harmful human activity.

At the moment, neither the Darwin Core nor ABCD standards have fields for recording the type of record-level metadata that is recommended here. A number of recommendations have been made to the Taxonomic Databases Working Group (TDWG) for the inclusion of extra fields to the [Geospatial Element Definitions Extension](#) to Darwin Core (TDWG 2005) and/or to the Darwin Core itself. The recommendations included those shown in the table on the next page.

The ‘*DataSensitiveComments*’ field here is perhaps equivalent to the ‘*Access Constraints*’ field in most dataset level metadata. The sort of information at the dataset level may include something like:

*“This dataset is only available to the public at a summary resolution for the following reason. Some of the information held within this dataset relates to species that are vulnerable to human disturbance or prejudice. Two species (*Adelanthus lindenbergianus*, *Athalamia hyaline*) are significantly vulnerable to collecting. The full detail of this sensitive information may be made available under licence to specific organisations and individuals that need to know to avoid harm to the environment. Please contact the provider for more information.”*

Until such time as these standards and protocols are modified, it is recommended that the data be documented in comment fields, and as far as possible to record the same type of information that would be included in the recommended fields above – i.e.

- That the data are sensitive;
- The primary reasons the data are regarded as sensitive (see Criteria 1-4 the Chapter on *Determining Sensitivity*) along with supporting rationale;

- The date that the sensitivity of the data should be reviewed;
- Precision of the data made available;
- Precision of the original data stored or retained.

Field	Comments
DataSensitiveIndicator	Y/N flag that the observation is sensitive.
DataSensitiveReason	The primary reason why the data are sensitive. Suggested format is either a picklist with values derived from Criteria 1-4 above (or a text field that combines the statements 1a-4g attached to those criteria).
DataSensitiveComments	Further information on the reason(s) or supporting rationale for determining relevance of the Criteria for this record as recommended above. [Free Text]
SensitiveDateForReview	A date field documenting when the sensitive nature of the date should be reviewed. Especially important where the sensitivity is just awaiting publication of results, etc.
PrecisionDataProvided	The scale or the precision of the data made available via the Darwin Core record – may be done as precision, e.g. <ul style="list-style-type: none"> • 0 = 1 degree • 1 = 0.1 degree • 2 = 0.01 degree • 3 = 0.001 degree • 4 = 0.0001 degree
PrecisionDataStored	The scale or the precision of the data made stored or retained by the data custodian – may be done as precision, e.g. <ul style="list-style-type: none"> • 0 = 1 degree • 1 = 0.1 degree • 2 = 0.01 degree • 3 = 0.001 degree • 4 = 0.0001 degree • Etc. or <p>may be more free text, such as ‘1 minute’, ‘0.1 minute’, ‘1 second’, etc. depending on how data are stored.</p>

Authentication and Authorisation

As recommended by the experts' workshop, and identified by many in the on-line survey, responsibility for determining who may or may not have access to detailed data on sensitive data, possibly through use of secure log-on, or one-off data license agreements, must be with the data providers.

It was also agreed at the workshop that it is not the role of GBIF to manage the identification, verification or authorisation of users, nor to control authentication or log-on at the Data Portal, but it may have a role in providing guidance and a suitable authentication method to the Nodes.

It was reported at the experts' workshop that the technical issues relating to the authentication of a group or individual, and the use of roles, etc. is not a difficult task. There are several well established protocols and working systems for authentication in use and these could easily be adapted for use by data providers.

The main issue is in determining who the authorized users should be and how to determine who are *bona-fide* users and who are not. This is a difficult issue and one that will need to be explored over time. It is not something that can be recommended in this best practices document; however the earlier report (Chapman 2007b) did make a number of recommendations on how this issue may be further explored.

It has been recommended that GBIF explore the issue of authentication with the view to providing appropriate mechanisms that help data providers. It is therefore recommended that data providers wishing to develop secure authentication for their databases discuss the issue with GBIF, or with their GBIF Node.

The recommendation made to GBIF in the earlier report (Chapman 2007b) was that:

GBIF explore the issue of authentication with the view to providing appropriate mechanisms that help data providers identify users who can dig deeper and how. Although GBIF shouldn't have a role (at this stage at least) in vetting users, or in placing controls on the GBIF Portal, it does have a role in providing guidance and assisting Nodes in implementing a suitable and robust authentication method.

References

- Chapman, A.D. 2006. *Questionnaire on Dealing with Sensitive Primary Species Occurrence Data – Summary of responses*. 61 pp. Copenhagen: GBIF.
http://www.gbif.org/prog/digit/sensitive_data/Summary_of_Responses_-_03.pdf. [Accessed 8 Apr. 2007].
- Chapman, A.D. 2007a. Workshop on Dealing with Sensitive Species Occurrence Data. Held at: NatureServe Offices, Arlington, Virginia, USA. 6-7 March 2007. Report. Copenhagen: GBIF. 30 pp. <http://www.gbif.org/>
- Chapman, A.D. 2007b. *Dealing with Sensitive Primary Species Occurrence Data. Report*. Report to the Global Biodiversity Information Facility 60pp. <http://www.gbif.org/>. Copenhagen: GBIF.
- Chapman, A.D. and Wiczorek, J. (eds). 2006. *Guide to Best Practices for Georeferencing*. BioGeomancer Consortium. Copenhagen: Global Biodiversity Information Facility. 90pp. ISBN: 87-92020-00-3.
<http://www.gbif.org/prog/digit/Georeferencing> **See also Chapter 5 in this Manual.**
- Department of Environment and Conservation – NSW. 2007. *Threatened Species Information Disclosure Policy* (Version 3 Amended March 2007).
http://www.nationalparks.nsw.gov.au/npws.nsf/content/sensitive_species_policy [Accessed 15 Mar 2007].
- Llinás, J.V. 2005. *Data and Information on Biodiversity and its Protection in the Digital Realm* Ver. 1. Bogotá, Colombia: Biological Resources Research Institute Alexandre von Humboldt. 43pp.
- National Biodiversity Network Trust. 2002. *NBN Data Exchange Principles*. Version 3.2, April 2002.
<<http://www.nbn.org.uk/downloads/files/DataExchange%20principles%202002.pdf>> [Accessed 27 Mar 2007].
- National Biodiversity Network Trust. 2004. *The ‘Environmental Exception’ and access to information on sensitive features*. Version 1.3.2, Countryside Agencies’ Open Information Network Environmental Information Regulations Guidance Note No. 1. Linked from www.nbn.org.uk/eir [Accessed 27 Mar 2007].
- TDWG. 2005. *Geospatial Extension to Darwin Core*. Taxonomic Databases Working Group.
<http://wiki.tdwg.org/twiki/bin/view/DarwinCore/GeospatialExtension> [Accessed 1 Apr 2007].

Appendix: Scenarios using Criteria 1 and 2 as Triggers

The following sets of scenarios show how the criteria statements given in the Chapter on *Determining Sensitivity* may be used to develop summary statements for documenting the reasons why a taxon may be regarded as sensitive. The summary statement (in the white boxes), should also include supporting rationale, such as specific types of harm, etc. For example in the second scenario (B) – the full statement may read something like:

“Taxa could be at risk from harm from diseases carried on the wheels of forestry machinery but occurrence is not affected by data availability.”

This may apply to a species of plant in a forestry area susceptible to *Phytophthora* attack, the fungi being transferred on the wheels of forestry vehicles.

Criterion 1:

Scenario A

Criterion statement(s)	Summary statement
1a – There is no significant risk of a harmful human activity.	The taxon is not sensitive.

Scenario B

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon could be at risk from harm but likelihood of harm is not affected by data availability.
1d – There is currently no established evidence of actual harm to the taxon.	
1f – Availability of biodiversity data will not increase the likelihood of the harmful human activity taking place.	

Scenario C

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon could be at risk from harm and the likelihood of harm is affected by data availability.
1d – There is currently no established evidence of actual harm to the taxon.	
1e – Availability of biodiversity data will increase the likelihood of the harmful human activity taking place.	

Scenario D

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon is at risk from harm and there is evidence to support this, but occurrence is not affected by data availability.
1c – There is established evidence of actual or recent harm to the taxon.	
1f – Availability of biodiversity data will not increase the likelihood of the harmful human activity taking place.	

Scenario E

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon is at risk from harm, there is evidence to support this, and occurrence is affected by data availability.
1c – There is established evidence of actual or recent harm to the taxon.	
1e – Availability of biodiversity data will increase the likelihood of the harmful human activity taking place.	

Criterion 2:

Scenario F

Criterion statement(s)	Summary statement
2b – The taxon is not significantly vulnerable to the harmful human activity.	The taxon is not significantly vulnerable to the harmful activity, and is not vulnerable to that activity over its total range and there are areas where the taxon is not at significant risk from that activity.
2d – The taxon is not vulnerable to harmful human activity over its total range and/or there are areas where the taxon is not at significant risk.	

Scenario G

Criterion statement(s)	Summary statement
2a – The taxon has characteristics that make it significantly vulnerable to the harmful human activity.	The taxon is significantly vulnerable to the harmful activity, but is not vulnerable to that activity over its total range and there are areas where the taxon is not at significant risk from that activity.
2d – The taxon is not vulnerable to harmful human activity over its total range and/or there are areas where the taxon is not at significant risk.	

Scenario H

Criterion statement(s)	Summary statement
2a – The taxon has characteristics that make it significantly vulnerable to the harmful human activity.	The taxon is significantly vulnerable to the harmful activity, and is vulnerable to that activity over its total range.
2c – The taxon is vulnerable to harmful human activity over its total range.	

Glossary

Authentication: — refers to the determination of a user's identity, as well as determining what a user is authorized to access. The most common form of authentication is user name and password, although this also provides the lowest level of security.

Authorisation: — refers to the process of determining which individuals can be afforded different access rights for authentication and data access.

Generalisation: — refers here to any modifications carried out to source data to conceal sensitive content, typically by reducing the precision of the data (such as reporting at the level of a watershed, grid or county, citing just the nearest named place, or by deleting some parts of the data). In geographic terms it refers to the conversion of a geographic representation to one with less resolution and less information content; traditionally associated with a change in scale. Also referred to as: *fuzzying*, *dummying-up*, etc.

Record-level Metadata: — refers to documentation at the level of a record rather than for a complete dataset. In this document it largely refers to documentation of the sensitivity status of the record (or the species of which it is a part) along with access constraints pertaining to the record and details of any generalisation of the data.

Sensitive data: — any data, that because of their nature, a data provider does not want to make available in their raw state, e.g. precise localities of endangered taxa.

Index to Chapter 6

- ABCD protocol, 13
- authentication, **18**
 - definition, **19**
- authorisation, **18**
 - definition, **19**
- Categories of Sensitivity, 4, **8**
 - determining, **7**
- collector's number, 11
- co-relational analysis, 11
- Criteria for Determining Sensitivity, **4**
 - scenarios, **23**
- Darwin Core, 13, 14, 16
- Data License Agreements, 2
- Data Sharing Agreements, 2
- DataSensitiveComments, **16**
- DataSensitiveIndicator, 16
- DataSensitiveReason, **16**
- date of collection, 11
- decimal degrees, 13
- decimal geographic grid, 13
- Department of Environment and Conservation (NSW), 4, 8
- documentation, **15**
 - of generalized georeferences, **13**
- dummying up, 19
- ECat, 10
- experts' workshop, 2
- fuzzying, 19
- GBIF
 - Data Portal, 2, 18
 - Nodes, 18
 - Secretariat, 2
- generalisation
 - definition, **19**
 - spatial information, **13**
 - textual information, **11**
- georeferences
 - generalizing, 13
- Geospatial Element Definitions Extension, 14, 16
- Guide to Best Practices in Georeferencing, 2
- habitat, 11
- harm
 - impact of, 4, **5**
 - risk of, 4, **5**
- impact of harm, 4, **5**
- landholder information, 11
- lists of sensitive taxa, **10**
- living persons
 - names of, 11
- locality information, 11
- maximum uncertainty
 - estimate, 16
- metadata, **15**
- National Biodiversity Network (NBN), 4
- on-line survey, 2
- PrecisionDataProvided, 16
- PrecisionDataStored, 16
- Principles, **3**
- randomisation
 - definition, **19**
- risk of harm, 4, **5**
- sensitive data
 - definition, **19**
- sensitive taxa
 - listing of, **10**
- SensitiveDateforReview, **16**
- sensitivity
 - Categories of, 4, **8**
 - determining, **4**
 - documenting, **15**
 - of data, 4
 - determining, **6**
 - sensitivity criteria, 4
 - spatial information
 - generalisation of, **13**
- Survey Monkey, 2
- taxonomic names, 11
- textual information, **11**

GBIF Glossary and Acronym Expansion

ABCD Schema	<p>The <i>Access to Biological Collection Data</i> (ABCD) Schema is the product of a joint TDWG and CODATA initiative to develop a standard for distributed data retrieval from collection data bases. The schema seeks to cover data exchange for all kingdoms and for both specimen and observation records.</p> <p>http://bgbm3.bgbm.fu-berlin.de/TDWG/CODATA/Schema/default.htm</p>
BDWorld	<p>Biodiversity World http://www.bdworld.org/</p>
Berlin Taxonomic Information Model	<p>The <i>Berlin Taxonomic Information Model</i> is a database model for handling the complexity of taxonomic names, in particular botanical names.</p> <p>http://www.bgbm.org/biodivinf/docs/bgbm-model/</p>
BioCAsE	<p>The <i>Biological Collection Access Service</i> Protocol is derived from the DiGIR protocol and supports web-based searches for XML data. It has been used in particular for data exchange using the ABCD schema. http://www.biocase.org/</p>
Biodiversity	<p><i>Biodiversity</i>, the short form of "Biological diversity," means the variability among living organisms from all sources including, <i>inter alia</i>, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are part; this includes diversity within species, between species and of ecosystems (Convention on Biological Diversity, Art. 2, paragraph 1).</p>
Biodiversity data	<p><i>Biodiversity data</i> refers to any data which presents information about the world's biodiversity, including <i>species/observation</i> data, <i>general resource</i> data and <i>name list</i> data.</p>
Biodiversity Database Interoperability	<p>The <i>Biodiversity Database Interoperability</i> segment of the <i>GBIF information architecture</i> manages the portal's access to <i>web services</i> on other machines and handles the complexities of issuing a data requests to multiple providers which may be using heterogeneous protocols and data standards to present their data.</p>
Biologia Centrali-Americana Project	<p>The <i>Biologia Centrali-Americana (BCA) Project</i> has goals which include the delivery of a digitised version of the 58 biological volumes of the <i>Biologia Centrali-Americana</i> (a fundamental resource on the Neotropical flora and fauna).</p> <p>http://www.sil.si.edu/BCAProject</p>
BioMOBY	<p>BioMOBY is an international research project involving biological data hosts, biological data service providers, and coders whose aim is to explore various methodologies for biological data representation, distribution, and discovery. http://www.biomoby.org</p>
BioNET International	<p>BioNET-INTERNATIONAL is dedicated to supporting sustainable development by helping developing countries to overcome the taxonomic impediment by becoming self-reliant in taxonomy. http://www.bionet-intl.org/</p>
CBD	<p><i>Convention on Biological Diversity</i>. The Convention's member countries regularly share ideas on best practices and policies for the conservation and sustainable use of biodiversity with an ecosystem approach. http://www.biodiv.org/default.shtml</p>
CBOL	<p><i>Consortium for the Barcode of Life</i> is an international initiative devoted to developing DNA barcoding as an accurate and reliable tool for scientific research on the taxonomy of plant and animal species. http://barcoding.si.edu/index_detail.htm</p>
CEPDEC	<p><i>Capacity Enhancement Plan for Developing Countries</i> of GBIF under Strategic Plan 2007 - 2011</p>
CGIAR	<p><i>Consultative Group on International Agricultural Research</i>. http://www.cgiar.org/</p>
CHM	<p><i>Clearing House Mechanism</i> of the Convention on Biological Diversity promotes the sharing of information and technologies for working with biodiversity.</p> <p>http://www.biodiv.org/chm/default.aspx</p>

CITES	<i>Convention on International Trade in Endangered Species of Fauna and Flora</i> aims to ensure that international trade in specimens of wild animals and plants does not threaten their survival. http://www.cites.org/
CMS	<i>Convention on the Conservation of Migratory Species</i> aims to conserve terrestrial, marine and avian migratory species throughout their range. http://www.cms.int/
CODATA	The <i>Committee on Data for Science and Technology (CODATA)</i> is one of the bodies working on the development of the <i>ABCD Schema</i> . http://www.codata.org/
Collection coden	An acronym or other abbreviation that identifies a particular collection, for example “GH” for Gray Herbarium of Harvard University
CoLp	The <i>Catalogue of Life partnership</i> includes <i>Species 2000</i> and <i>ITIS</i> and is a grouping of organisations with the common goal of producing a unified catalogue of the names of all organisms.
CRIA	The <i>Centro de Referência em Informação Ambiental</i> (Reference Center on Environmental Information) is a not-for-profit, non-government organization. Its aim is to contribute towards a more sustainable use of Brazil's biodiversity through the dissemination of high quality information and education. http://www.cria.org.br/
Data Provider	A <i>Data Provider</i> is any computer within the <i>GBIF Network</i> that offers <i>data services</i> to the rest of the network. This term is also used to refer to the persons, institutions or organisations that share data.
Data Provider Toolkit	The <i>Data Provider Toolkit</i> will be developed by the <i>GBIF Secretariat</i> as a set of reusable components which may assist in the development of <i>Data Providers</i> .
D-Grid	Initiative zur Förderung eines Grid-basierten e-Science-Frameworks in Deutschland http://www.d-grid.de/
DiGIR	The <i>Distributed Generic Information Retrieval (DiGIR)</i> protocol is a development activity of the TDWG Access to Biological Collection Data (ABCD) subgroup. It is intended to support retrieval of structured data from multiple, heterogeneous databases. Both requests and replies are modeled as XML queries. It is currently being used to exchange data in the DwC format, but several groups are investigating use of DiGIR with the ABCD Schema. http://digir.sourceforge.net/
DNA	<i>Deoxyribonucleic Acid</i> a long linear polymer found in the nucleus of a cell and formed from nucleotides and shaped like a double helix; associated with the transmission of genetic information
DRM	<i>Digital Rights Management</i> is an umbrella term referring to any of several technical methods used to control or restrict the use of digital media content on electronic devices with such technologies installed.
DwC	The <i>Darwin Core</i> is a profile describing the minimum set of standards for search and retrieval of natural history collections and observation databases. It includes only core data elements which are likely to be available for the vast majority of specimen and observation records. http://tsadev.speciesanalyst.net/DarwinCore/darwin_core.asp
EDIT	<i>European Distributed Institute of Taxonomy</i> is an EC project to harness the power of the many taxonomic institutions in Europe by networking them to work synergistically
FAO	<i>Food and Agriculture Organisation</i> of the United Nations leads international efforts to defeat hunger. http://www.fao.org/UNFAO/about/index_en.html
Feedback	The <i>GBIF Network</i> allows users of its <i>data services</i> to provide feedback to the data providers. This function is implemented as the <i>User Feedback Service</i> , a <i>web service</i> offering an interface to pass a text message to the provider of any data item. This message is transmitted to the relevant <i>Data Provider</i> administrator as an e-mail.
GBIF	The <i>Global Biodiversity Information Facility</i> is an international organisation with the goal of making the world's <i>biodiversity data</i> freely and universally available. Its members include a wide range of countries and international organisations, and the GBIF Secretariat is based in Copenhagen, Denmark. http://www.gbif.org/

GBIF Communications Portal	The <i>GBIF Communications Portal</i> is a community resource that provides news, articles, events, documents and other linkages of use to the GBIF community. http://www.gbif.org/
GBIF Data Portal	The <i>GBIF Data Portal</i> is the node (or set of replicated nodes) which provides a number of key services to the rest of the <i>GBIF Network</i> to support access to the data distributed through the network. These services include the management of the <i>Registry</i> , the <i>Index</i> , access to the electronic catalogue of names of species, and a set of search tools. The <i>GBIF Data Portal</i> is not itself a <i>Data Provider</i> but serves as an integration point for all data in the network. http://www.gbif.net/ See also: http://wiki.gbif.org/dadiwiki/wikka.php?wakka=HomePage
GBIF Network	The <i>GBIF Network</i> is the entire network of computers and networks which comes together to provide the common pool of biodiversity data which <i>GBIF</i> presents.
GBIF Participant	A country or international organisation or economy that signs the MoU and agrees to carry out the activities indicated therein.
GenBank	<i>GenBank</i> is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. http://www.psc.edu/general/software/packages/genbank/genbank.html
Geographic service data	In the future <i>GBIF</i> will require access to an increasing number of external <i>data services</i> in areas related to biodiversity. One of the key areas will be <i>geographic service data</i> , providing information such as gazetteers or mapping services.
GIS	<i>Geographic Information Systems (GIS)</i> are computer systems and software which allow the combination of multiple data layers, each providing information about some characteristics of a geographic area. Such systems provide tools for mapping and analysing the data.
GISIN	<i>Global Invasive Species Information Network</i> http://www.gisinet.org
GISP	<i>Global Invasive Species Programme</i> was established in 1997 to address global threats caused by Invasive Alien Species (IAS), and to provide support to the implementation of Article 8(h) of the CBD.
GPP	<i>Global Pollination Project</i> (full title: Conservation and Management of Pollinators for Sustainable Agriculture, Through an Ecosystem Approach) of UNEP/GEF, executed by FAO.
Grid	<i>Grid</i> computing seeks to create new application models which can exploit large-scale pools of computing resources. http://www.gridforum.org/
GSPC	<i>Global Strategy for Plant Conservation</i> is a program of the CBD with the objective of halting the current and continuing loss of plant diversity. http://www.biodiv.org/programmes/cross-cutting/plant/default.asp
GTI	<i>Global Taxonomy Initiative</i> is a program of the CBD with the objective of addressing the lack of taxonomic information and expertise available in many parts of the world. http://www.biodiv.org/programmes/cross-cutting/taxonomy/
GUID	A <i>Globally Unique Identifier</i> is a numeric or text identifier which is guaranteed to be unique even at the global level used to improve connections between related data items even when served by different data providers.
HTML	<i>HyperText Markup Language</i> is the formatting language used to format most human-readable data on the Internet. http://www.w3.org/MarkUp/
HTTP	<i>Hyper Text Transfer Protocol</i> is a stable Internet specification/standard used globally. http://www.w3.org/Protocols/
IABIN	<i>Inter-American Biodiversity Information Network</i> is a regional initiative and is a Participant in GBIF. http://www.iabin.net/english/index.shtml
ICT	<i>Information and Communications Technologies</i>

Index	The <i>Index</i> is a component of the <i>GBIF information architecture</i> . It uses the service <i>metadata</i> held in the <i>Registry</i> to access all <i>data services</i> connected to the <i>GBIF Network</i> and to generate a central index for accessing <i>biodiversity data</i> .
Informatics	The use of <i>ICT</i> to make data and information automatically available via the Internet.
Interoperability	The ability of systems to provide services to and accept services from other systems and to use the services so exchanged to enable them to operate effectively together. With respect to software, the term interoperability is also used to describe the capability of different programs to read and write the same file formats and utilise the same protocols.
IPGRI	<i>International Plant Genetic Resources Institute</i> is the world's largest nonprofit agricultural research and training organisation devoted solely to the study and promotion of agricultural biodiversity. http://www.ipgri.cgiar.org/index.htm
IPI	<i>International Initiative for the Conservation and Sustainable Use of Pollinators</i> is an issue within the Agricultural Biodiversity Work Programme of the CBD. http://www.biodiv.org/programmes/areas/agro/pollinators.aspx
IPR	<i>Intellectual Property Rights</i> - In law, particularly in common law jurisdictions, intellectual property or <i>IP</i> refers to a legal entitlement which sometimes attaches to the expressed form of an idea, or to some other intangible subject matter.
ITIS	The <i>Integrated Taxonomic Information System (ITIS)</i> is a collaborative development between the United States of America, Canada and Mexico to develop catalogues of species names. http://www.itis.usda.gov/ , http://www.agr.gc.ca/itis and http://siit.conabio.gob.mx .
IUCN	The World Conservation Union, formerly known as the <i>International Union for the Conservation of Nature</i> http://www.iucn.org/
LDAP	<i>LDAP (Lightweight Directory Access Protocol)</i> is a model and protocol for storing and retrieving hierarchically-arranged information. It is designed to run directly over TCP/IP.
MAB	UNESCO's Programme on <i>Man and the Biosphere</i> develops the basis, within the natural and the social sciences, for the sustainable use and conservation of biological diversity, and for the improvement of the relationship between people and their environment globally. http://www.unesco.org/mab/about.htm
Metadata	<i>Metadata</i> are data records that provide descriptive information about other data. In the context of <i>GBIF</i> , <i>metadata</i> provides information about the suppliers of <i>biodiversity data</i> and about the origins and purpose of those data.
Mirror site	A <i>mirror site</i> is a replica of a central site, established to protect data from hardware failure, allow faster downloads, and to balance demand load. http://en.wikipedia.org/wiki/Mirror_site
MoC	<i>Memorandum of Cooperation</i>
MoU	<i>Memorandum of Understanding</i>
MyGrid	The <i>myGrid</i> project aims to exploit the growing interest in Grid technology, with an emphasis on the Information Grid, and provide middleware layers that make it appropriate for the immediate needs of bioinformatics. It is a UK e-Science project funded by the EPSRC involving five UK universities, the European Bioinformatics Institute and many industrial collaborators. http://www.mygrid.org.uk/index.php?&MMN_position=1:1&MMN_position=1:1
Name data	<i>Name data</i> refers to <i>structured data</i> providing information about taxonomic names and their relationships.
NCBI	The U.S. <i>National Center for Biotechnology Information</i> , part of the National Library of Medicine, in turn part of the National Institutes of Health. http://www.ncbi.nlm.nih.gov/
NCEAS	The U.S. <i>National Center for Ecological Analysis and Synthesis</i> http://www.nceas.ucsb.edu

NESCent	The U.S. <i>National Evolutionary Synthesis Center</i> http://www.nescent.org/main/
NODES	The GBIF committee comprising the managers of all <i>Participant Nodes</i>
Nomenclator	A <i>nomenclator</i> is a listing of the scientific names of a group of organisms, such as <i>Nomenclator Zoologicus</i> (http://uiio.mbl.edu/NomenclatorZoologicus/) or <i>Nomenclator Ipomoeae</i> (http://www.fau.edu/divdept/biology/people/daustin/nomen-1.htm)
OBIS	<i>Ocean Biogeographic Information System</i> , a component of the Census of Marine Life (CoML), provides an important component of GBIF data – those from the marine realm. http://www.iobis.org/Welcome.htm
OECD	<i>Organisation for Economic Cooperation and Development</i> http://www.oecd.org
OGC	<i>Open Geospatial Consortium</i> The Open Geospatial Consortium, Inc. (OGC) is a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services. http://www.opengeospatial.org/
Ontology	A formal ontology is a controlled vocabulary expressed in an ontology representation language. This language has a grammar for using vocabulary terms to express something meaningful within a specified domain of interest. However, the term is also used to refer to several different things: glossaries, data dictionaries, thesauri. http://www.metamodel.com/article.php?story=20030115211223271
Participant Node	A <i>Participant Node</i> is node within the <i>GBIF Network</i> established by a GBIF Participant as its contribution to the <i>GBIF Network</i> . It may also act as a <i>Data Provider</i> and is likely to service as a central registration point and access point for a number of other <i>Data Providers</i> .
Participant Node Toolkit	The <i>Participant Node Toolkit</i> will be developed by the <i>GBIF Secretariat</i> as a set of reusable components which may assist <i>GBIF</i> participant countries and organisations to develop their <i>Participant Nodes</i> .
Portal	<i>Portal</i> is used as a general term to refer to a web site that offers a single access point for users to retrieve content from a wide variety of sources.
Portal Toolkit	The <i>Portal Toolkit</i> is a toolkit offered by the <i>GBIF Secretariat</i> to assist <i>Participant Nodes</i> to develop their own <i>Support Services</i> components. It is based on the Zope web server and supports content syndication. http://circa.gbif.net/Members/irc/gbif/ict/library?!=/download_gbif_tools
Ramsar Convention	The Convention on Wetlands, signed in Ramsar, Iran, in 1971, is an intergovernmental treaty which provides the framework for national action and international cooperation for the conservation and wise use of wetlands and their resources. http://www.ramsar.org/
Registry	The <i>Registry</i> is a component of the <i>GBIF information architecture</i> , which may also be redeployed within <i>Participant Nodes</i> . It is responsible for maintaining <i>metadata</i> about <i>Data Providers</i> and <i>web services</i> .
SDD	The <i>TDWG Structured Descriptive Data</i> subgroup has the task of developing an interoperability standard for descriptive data providing information about character states for different organisms. http://www.tdwg.org/sddhome.html
SEEK	<i>Science Environment for Ecological Knowledge</i> is a five year initiative designed to create cyberinfrastructure for ecological, environmental, and biodiversity research. http://seek.ecoinformatics.org/
SIS	<i>Species Information Service</i> , a world-wide species information resource on the status and distribution of species threatened with extinction. http://www.iucn.org/themes/ssc/programs/sisindex.htm
Species 2000	<i>Species 2000</i> is an international organisation with the goal of enumerating all known species of plants, animals, fungi and microbes on Earth as the baseline dataset for studies of global biodiversity. http://www.sp2000.org/

Species Analyst	<i>The Species Analyst</i> is a research project developing standards and software tools for access to the world's natural history collection and observation databases. The Species Analyst is based at the University of Kansas Natural History Museum and Biodiversity Research Center. http://tsadev.speciesanalyst.net/
Specimen / observation data	Throughout this document <i>specimen/observation data</i> refers to data describing individual specimens or observations of organisms identified by taxon.
Data (or metadata) standard	Technical <i>standards</i> define a set of properties that a product or service should have. Standards are laid down by an organisation, such as TDWG and GBIF, that brings together representatives of producers and users of the type of product or service to establish the standard(s) in question.
Structured data	Throughout this document <i>structured data</i> refers to any data for which the structure and permitted content have been clearly defined.
TAPIR	<i>TDWG Access Protocol for Information Retrieval</i> is the “next generation” protocol which combines the capabilities of BioCAsE and DiGIR
TaXMLit	<i>Taxonomic XML literature</i> protocol developed by the Biologia Centrali-Americana project to enable exchange of digital literature.
TCP/IP	<i>Transmission Control Protocol/Internet Protocol</i> is the main transport protocol used within the Internet to allow any two machines to communicate.
TCS	<i>Taxon Concept Schema</i> developed by TDWG with support from GBIF and SEEK to enable the efficient handling of names and taxonomic data within the GBIF information architecture. http://tdwg.napier.ac.uk/TCS_1.0/docs/publications.html
TDWG	The <i>Taxonomic Database Working Group (TDWG)</i> is an international body established to define standards for use in biological data projects. http://www.tdwg.org/
Thesaurus	A <i>thesaurus</i> is a networked collection of controlled vocabulary terms. http://www.metamodel.com/article.php?story=20030115211223271
uBIO	<i>Universal Biological Indexer and Organiser</i> is an initiative within the science library community to partner with other international efforts such as GBIF to create and utilize a comprehensive and collaborative catalog of known names of all living (and once-living) organisms. http://www.ubio.org/
UDDI	<i>Universal Description, Discovery and Integration (UDDI)</i> is a registry technology and protocol available for use in creating web-based registries of <i>web service</i> implementations. http://www.uddi.com/
UNEP	<i>United Nations Environment Programme</i> established to provide leadership and encourage partnership in caring for the environment by inspiring, informing, and enabling nations and peoples to improve their quality of life without compromising that of future generations. http://www.unep.org/Documents.Multilingual/Default.asp?DocumentID=43
UNESCO	United Nations Educational, Scientific and Cultural Organisation encourages international peace and universal respect by promoting collaboration among nations. http://www.unesco.org
Unstructured data	<i>Unstructured data</i> refers to any data for which the structure and permitted content are not clearly defined. Such data may in fact be formatted according to a definite structure, but this structure is unknown to the system processing the data.
WCMC	UNEP <i>World Conservation Monitoring Center</i> provides information for policy and action to conserve the living world. http://www.unep-wcmc.org/
Web	The World Wide Web ("WWW", "W3", or simply "Web") is an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URIs). The term is often mistakenly used as a synonym for the Internet, but the Web is actually a service that operates <i>over</i> the Internet. http://en.wikipedia.org/wiki/World_Wide_Web

Web Service	A <i>web service</i> is any computing service which is published and accessible across the Internet and offers a standardised <i>XML</i> interface allowing users to invoke its function. Most of the <i>web services</i> discussed in this document provide access to biodiversity data.
Web-enabled Taxonomy	Any implementation of Internet technology that assists and enhances the development of the classification, nomenclature and taxonomy of groups of organisms
WFS	A <i>Web Feature Service</i> allows a client to perform data manipulation operations on a set of geographic features
WMS	A <i>Web Mapping Service</i> allows a client to generate a map online in real time
XML	The <i>eXtensible Markup Language (XML)</i> is a simple document format developed from SGML to support electronic publishing. It provides a flexible model particularly for structuring textual data. The <i>XML</i> language is supplemented by technologies such as <i>XML Schema</i> , <i>XPath</i> and <i>XQuery</i> to produce a powerful model for processing electronic data. http://www.w3.org/XML/

The GBIF Data portal

... a practical "hands-on" tutorial

<http://data.gbif.org>

GLOBAL BIODIVERSITY INFORMATION FACILITY



The GBIF Data Portal: A practical, “hands-on” tutorial

Table of Contents

Introduction	1
Scientific names and classification in the GBIF portal	2
Maps in the GBIF portal	4
Finding information about a species or group of organisms	6
Searching for a species or group	6
Example: Desert Kangaroo Rat	6
Browsing the classification	8
Example: Komodo Dragon	9
The species or group Overview page	10
Finding information about a country	12
Searching for species that occur within a country	12
Browsing the country list	12
The country Overview page	13
Example: Ecuador plant species	14
Example: Exploring occurrences recorded within Indonesia	15
Searching for Occurrences	16
Using search filters	16
Available filter categories	17
Viewing details of an Occurrence record	18
Finding information about a dataset	20
Searching for a dataset	20
Browsing the dataset list	20
The dataset Overview page	21
Example: Finding datasets that have georeferenced records	22
Example: Finding data providers from Denmark	23
Example: Migratory birds	24
Examples: Viewing GBIF occurrence data in Google Earth	26
Species of <i>Mimosa</i>	26
Hummingbird record density	28

Introduction

The GBIF data portal is a service that provides access to millions of scientific data records that are being shared via the GBIF network. These data are generously made available by a wide range of institutions and organisations from around the world. To see the range of data providers involved, please see the list of data providers and datasets_ (data.gbif.org/dataset/).

The two types of data currently being shared through the GBIF Network are:

- Species occurrence records (based on specimens and observations) - information about the occurrence of species at particular times and places.
- Names and classifications of organisms - information on the names (both scientific and common) used for species and on the classification of those organisms into taxonomic hierarchies.

GBIF does not use just one taxonomic classification. And, there is not a complete electronic catalogue of all the scientific names that have ever been published available for GBIF to use. Nonetheless, classification schemes and lists of scientific names are essential to searching the occurrence data. For help in understanding how the GBIF portal addresses this problem, see [Scientific names and classification in the GBIF portal](#), page 2.

The power of the data shared by the GBIF network is that much of it can be mapped geospatially, which in turn makes it amenable to a vast array of analyses and therefore useful to many sectors of society. [Maps in the GBIF portal](#), page 4, explains the characteristics and use of maps in portal search results.

The portal's search function is a sophisticated tool that allows a user to rapidly and efficiently look for and find data records of interest from among millions of records made available via the GBIF network. The data can be sorted along taxonomic lines, by geography, or by timeframe. Through the application of a number of possible filters, records

Note regarding display of the GBIF portal:

- The standard layout for the portal has been optimised for screen sizes of 1024 by 768 or larger. If you are using a smaller display, we recommend that you visit the portal [Settings](#) page and select the look-and-feel for smaller displays.
- The portal is best viewed using the most recent version of Firefox (www.mozilla.com), Opera (www.opera.com), or Internet Explorer (www.microsoft.com), or Safari (www.apple.com).



free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

Settings

Look and Feel

Large display (recommended for displays 1024x768 and larger)

Smaller display (recommended for displays 800x600 and larger)

that match combinations of geography, time and taxonomy criteria can be isolated for further study (see [Searching for Occurrences](#), page 16).

This tutorial does not exhaust all the possible search capabilities of the portal, but it does provide an introduction to the main features, and provides some examples for combining search parameters to focus on records of interest.

If you are reading the hard copy or using this tutorial online, you can try the step by step instructions in (another) browser window as you go along. If you are using the CD version, the given examples will work but you will need an Internet connection to explore data through the portal itself.

Scientific names and classification in the GBIF portal

It is the aim of the GBIF data portal to guide users to relevant information on particular species and groups of organisms. To make this possible, it relies on classification hierarchies shared by its data providers.

Every dataset within the portal includes information on the names and classifications of the organisms included. Each species occurrence record includes a scientific name to which the organism was identified, and usually also includes some higher classification - sometimes the full set of species, genus, family, order, class, phylum and kingdom, but in other cases only a small subset of this information.

The scientific name for an organism may differ depending on the classification used. In particular, scientific names may change as scientists improve our understanding of the group concerned and species are moved into a different genus, or split into multiple species, or even combined into a single species. All of these changes can lead to a number of different names being in use for the same set of organisms. These different names for the same organism are called synonyms.

A further variation arises because some datasets are organised around a standard modern classification for the groups concerned (which means that the classification is consistent for all records within the dataset), but other datasets simply report the information that was recorded when the specimen was collected or the observation made. In this case, the same organism may appear under different names or in different classifications even within a single dataset.

This means that records for a particular species could be named and/or classified in different ways in different datasets. To address this problem, the portal makes use of authoritative classifications wherever possible. It uses these classifications to construct a working structure for the data and seeks to organise other names and classifications into the best place within this structure.

One major source of such authoritative classifications is the Catalogue of Life Annual Checklist (data.gbif.org/dataset/provider/218) which brings together scientifically reviewed checklists of different groups of organisms into one overall structure; it currently includes over one million species. This checklist includes synonyms and common names for many of the species it covers. The GBIF portal employs this checklist as the core of its own working structure and uses the synonym information to combine records where appropriate.

In addition, the portal uses the International Plant Names Index (data.gbif.org/dataset/provider/3) as a source for authoritative information on the names of plants, and the Index Fungorum (www.speciesfungorum.org/Names/Names.asp) as a source for authoritative information on the names of fungi.

Free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

Search

HOME SPECIES COUNTRIES

Classification
according to: The Global Biodiversity Information Facility: GBIF Data Portal Classification (based on Catalogue of Life from specimen and observation data resources)

Find a scientific name within this classification

Jump to a name Search

All Kingdoms

- Kingdom: Animalia — see overview page
- Kingdom: Archaea — see overview page
- Kingdom: Bacteria — see overview page
- Kingdom: Chromista — see overview page
- Kingdom: Fungi — see overview page
- Kingdom: Plantae — see overview page
- Kingdom: Protozoa — see overview page
- Kingdom: Viruses — see overview page

www.gbif.org design © GBIF. Data providers retain all rights to data.

Download: Darwin Core records
Send: Feedback to TH



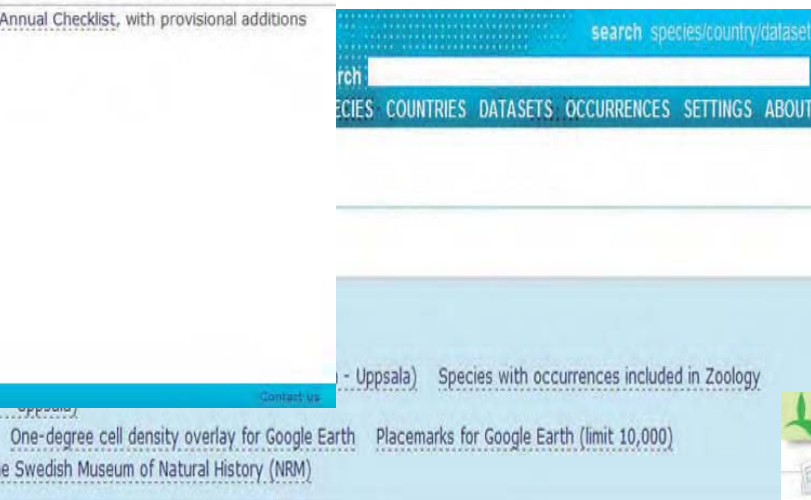
These resources cover a very large proportion of the species represented in GBIF data. However, there are still many records that bear names that are not included in these resources. The portal therefore uses an automated process to try to find the best placement within the hierarchy for each name that does not appear in these indices.

GBIF recognises that this process is error-prone, and that the results can at best only be tentative, but they do help to increase the probability that users will find records of relevance to their needs, even if the scientific names associated with those records are not included in the Catalogue of Life, the International Plant Names Index or the Index Fungorum.

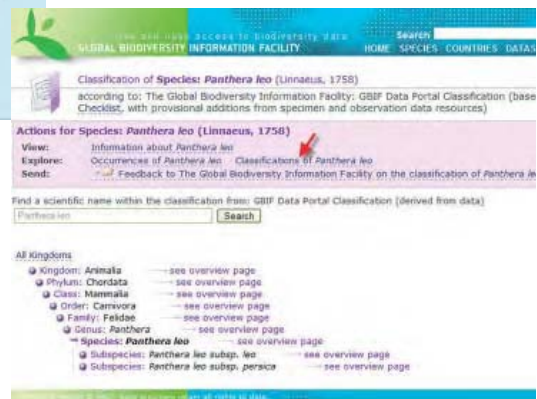
Names that have been inserted into the classification through this automated process are shown in grey in the Classification Browser and are included in the classification tree under the heading “unconfirmed names”.

The Classification Browser allows users to navigate this automatically-generated working structure. It is also possible to explore the classification used in each individual data resource. To do this, visit the Overview Page for the dataset and select the “Names and classification” link to the right of Explore in the blue Actions box at the top of the page. This will open

the Classification Browser to show just the names and classification used in that particular dataset.



The Classification Browser also allows a comparison among classifications for the current species or group of organisms across all datasets included in the portal. Select the link “Classifications of” in the pink Actions box at the top of the page.



This link opens a tabular view that shows the classification for the current species or group of organisms in all datasets included in the portal. Click on “view” at the right side of the table to see the whole classification employed within a particular dataset.

Taxon	Author	Accepted name	Rank	Dataset	Kingdom	Phylum	Class	Order	Family	Genus	Species	
<i>Panthera leo</i>	(Linnaeus, 1758)	Y	Species	Natural History Museum Rotterdam (NMR)	Animalia	Chordata	Mammalia	Carnivora	Felidae	<i>Panthera</i>	<i>Panthera leo</i>	View
<i>Panthera leo</i>		Y	Species	Animal Sound Archive Berlin								View
<i>Panthera leo</i>		Y	Species	Mammals (NRM)	Animalia	Chordata	Mammalia	Carnivora	Felidae	<i>Panthera</i>	<i>Panthera leo</i>	View
<i>Panthera leo</i>		Y	Species	Zoology (Museum of Evolution - Uppsala)	Animalia				Felidae	<i>Panthera</i>	<i>Panthera leo</i>	View
<i>Panthera leo</i>		Y	Species	Vertebrate specimens	Animalia	Chordata	Mammalia	Carnivora	Felidae	<i>Panthera</i>	<i>Panthera leo</i>	View
<i>Panthera leo</i>		Y	Species	Animal observations	Animalia		Mammalia				<i>Panthera leo</i>	View

Maps in the GBIF portal



The GBIF portal offers overview maps for the distribution of occurrence records for each species or group of organisms, for each country and for each data provider, dataset or data network. Similar maps are also offered for viewing the results of occurrence searches.

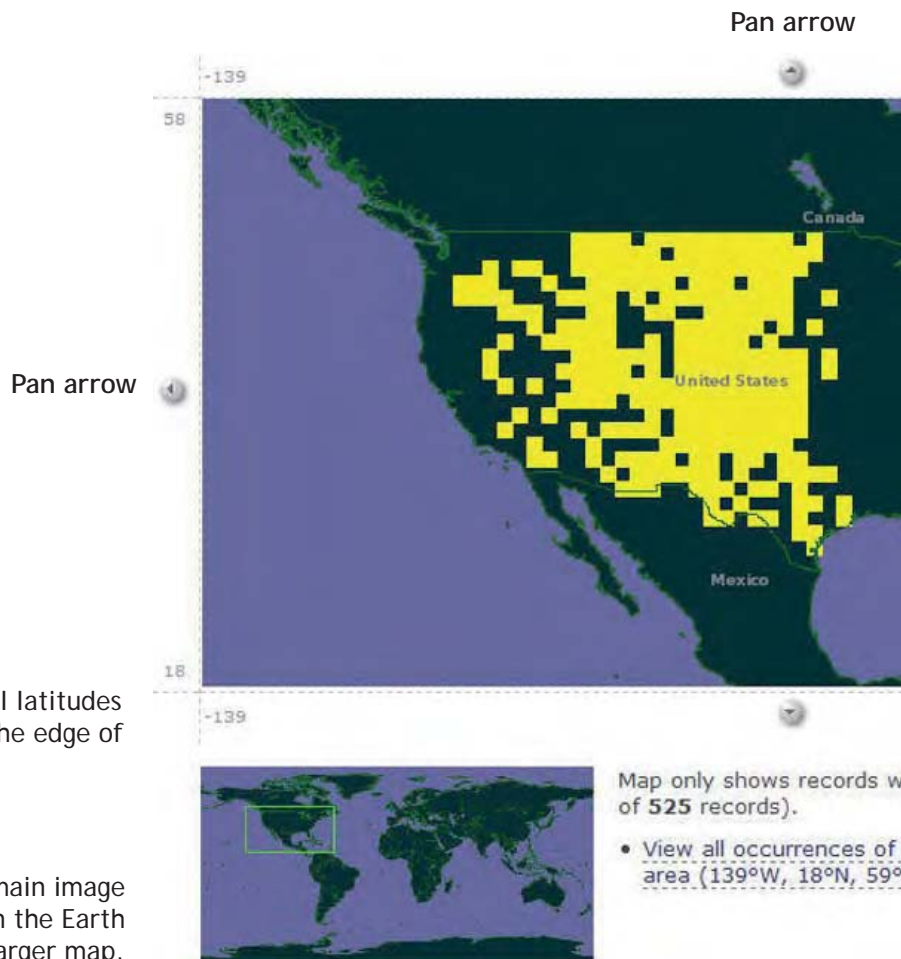
Only georeferenced records are shown on maps, and records with coordinates that are not consistent with the country named in the same record are also excluded.

The maps show the number of records that occur in each 1.0 degree by 1.0 degree cell on the globe, and in some cases in each 0.1 degree by 0.1 degree cell. GBIF Data Portal maps are intended as a first indication of distributions, and are not resolved further than 0.1 degree in order to keep searches rapid. Greater geospatial resolution can be obtained by exporting the data to other computer applications.

The maps generated within the GBIF portal show relatively little detail on underlying geography. Such underlying detail can be obtained by exporting the data to Google Earth (see the examples for hummingbirds, page 26 or plant species, page 27), or by downloading the data and mapping it against GIS layers.

Step-by-step guide

1. Navigate to any page showing an overview map, for example the overview page for any species (see Finding information about a species, page 6).

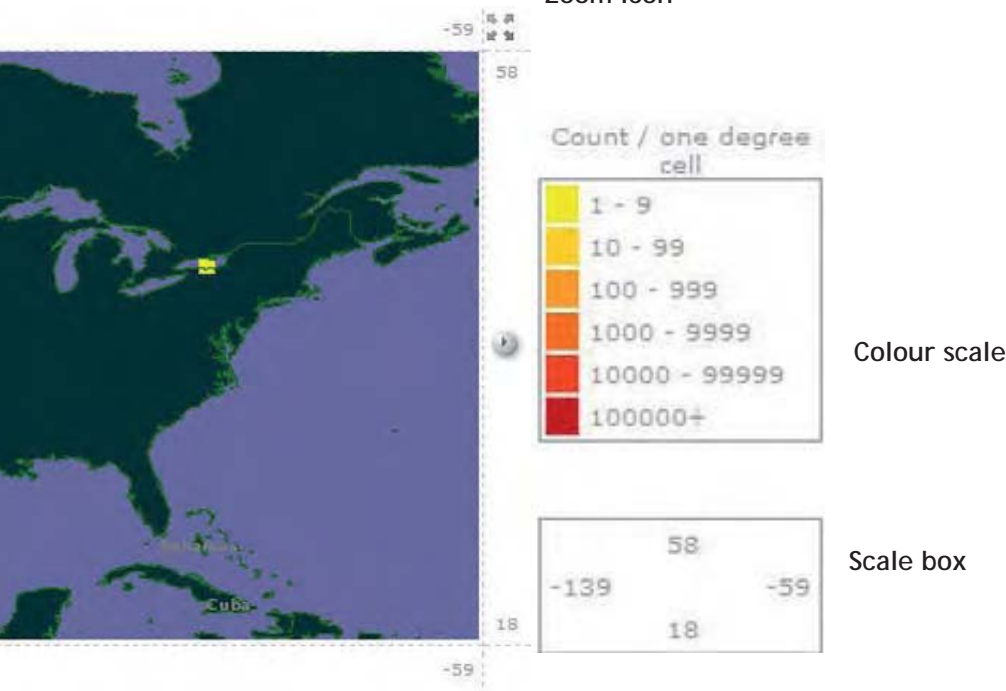


2. Each map shows decimal latitudes and longitudes around the edge of the image.

A small map below the main image indicates the area on the Earth depicted in the larger map.

- Each map includes a colour scale that indicates the meaning of the different colours on the map. These colours have the same significance on all maps; darker shades indicate increasingly larger numbers of records.
- Depending on the type of search, the map may be a view of the entire world, or it may be focused on a particular region if the relevant records are restricted to a smaller area (e.g. if the map is of the records from a particular country or of a species with a restricted distribution).
- Maps always show 1.0 degree cells apart from the highest resolution view, which shows 0.1 degree cells for an area spanning two degrees of longitude and one degree of latitude. The scale box (below the colour scale, to the right of the map) indicates the size of cells on each map.
- Moving the mouse over the map causes a red rectangle to appear. Move the mouse to move the rectangle. Clicking the mouse will cause the map to zoom in on the area marked by the red box (if the map is currently showing one degree cells) or to open the occurrence search view to see the records occurring within a 0.1 degree cell if the map is at high resolution.
- Click on the icon with four arrows at the top right corner of the map to zoom out again.
- Click on any of the four single arrows along the edges of the map to pan the view in that direction.

Zoom icon



with coordinates (482 records from a total

Gutierrezia sarothrae within the viewed
W, 58°N)

- In some cases there is a "View all" link below the map to open the occurrence search view to see georeferenced records that occur within the area currently shown on the map. In other cases, there is a count of the georeferenced records as well as a count of all records (both with and without coordinates), and some cases (as in this example), both.

Finding information about a species or group of organisms

The GBIF Portal provides access to data on the distribution of species in the form of species occurrence records (the details of an occurrence of a particular species at a particular place at a particular time), as well as the name(s) and classification(s) of each species, and links to additional information, if this is available. Because classification(s) are included, you can use them to search for data on a group of organisms that includes one or more species.

There are two ways to find information on a species or group: Search for it directly, or use the Classification Browser.

Searching for a species or group

The simplest way to find information on a species or group is to search for its name using the search box that is included on every page of the portal.

Step-by-step guide

1. From the home page of the portal or any other page, enter a scientific name or a common name in the Search box and click Search (or hit Enter/Return on your keyboard).

Example: Desert Kangaroo Rat

Find the scientific name and associated occurrence records

1. Begin by entering "kangaroo rat" into the search box on the Data Portal (data.gbif.org).

2. The search results page shows the first ten common name matches for "kangaroo rat." The "View all ..." link below this list leads to a page with all common name matches for "kangaroo rat." Common name information is given in parentheses after the scientific name and taxonomic classification information is given to the right of each link.

Search Results for: kangaroo rat

[Scientific names](#) [Common names](#) [Countries](#) [Datasets](#)

Scientific names

No scientific names matching "kangaroo rat"

Common names

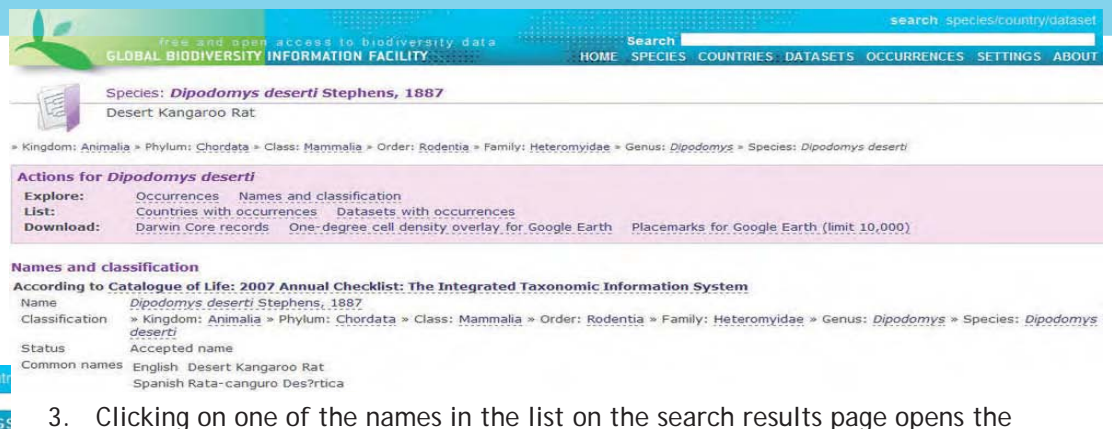
Species	Dipodomys agilis (English: Agile Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys californicus (English: California Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys compactus (English: Gulf Coast Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys deserti (English: Desert Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys elator (English: Texas Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys heermanni (English: Heermann's Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys ingens (English: Giant Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys merriami (English: Merriam's Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys microps (English: Chisel-toothed Kangaroo Rat)	Animalia - Chordata
Species	Dipodomys nitratoides (English: Fresno Kangaroo Rat)	Animalia - Chordata

[View all common names matching "kangaroo rat"](#)

Countries

No countries with names matching "kangaroo rat"

2. After you click "Agree" at the bottom of the Data Use Agreement (this will not appear again), the portal will return lists of names that match the search string, as well as names that include the search string. Scientific names are returned with an indication of their place in the overall classification to make it easier to select between multiple matches.
3. The search results are grouped into four categories (you may need to scroll down to see all the results, or click on the category headings near the top of the page):
 - a. Scientific Names
 - b. Common Names
 - c. Countries
 - d. Datasets
4. If the search string matches a large number of items in one of these categories, only the first 10 will be displayed. To see the rest of the matches, click on the link that begins "View all ..." at the end of the list, and then select from the full list (which may spread over multiple pages).
5. Click on a name under either of the categories Scientific Name or Common Name to see the Overview Page (explained below) for the chosen species or group of organisms.



free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

Search search species/country/dataset

HOME SPECIES COUNTRIES DATASETS OCCURRENCES SETTINGS ABOUT

Species: *Dipodomys deserti* Stephens, 1887
Desert Kangaroo Rat

Kingdom: Animalia » Phylum: Chordata » Class: Mammalia » Order: Rodentia » Family: Heteromyidae » Genus: *Dipodomys* » Species: *Dipodomys deserti*

Actions for *Dipodomys deserti*

Explore: Occurrences Names and classification
List: Countries with occurrences Datasets with occurrences
Download: Darwin Core records One-degree cell density overlay for Google Earth Placemarks for Google Earth (limit 10,000)

Names and classification

According to Catalogue of Life: 2007 Annual Checklist: The Integrated Taxonomic Information System

Name: *Dipodomys deserti* Stephens, 1887
 Classification: Kingdom: Animalia » Phylum: Chordata » Class: Mammalia » Order: Rodentia » Family: Heteromyidae » Genus: *Dipodomys* » Species: *Dipodomys deserti*
 Status: Accepted name
 Common names: English: Desert Kangaroo Rat
 Spanish: Rata-canguro Des?rtica

3. Clicking on one of the names in the list on the search results page opens the Overview page for that species. The Actions box provides links to additional information. Below the Actions Box is a listing of the classification(s) that are available in the Portal for the selected species.

4. The map on this Overview Page supplies a visual summary of georeferenced occurrence records for the Desert Kangaroo Rat. Note that below the map, you are told how many records there are that are mapped, and how many records there are in total.



Finding information about a species or group of organisms

Browsing the classification

To locate a species or group by browsing the full classification of organisms included in the portal through the Classification Browser, select the Explore Species link on the home page of the portal, or on the SPECIES link included in the banner at the top of every page.



Please note:

The classification shown through this view is an *automatically generated extension* of data that have been provided by taxonomic authorities, in particular the Catalogue of Life Annual Checklist (www.catalogueoflife.org/), the International Plant Names Index (www.ipni.org) and the Index Fungorum (www.speciesfungorum.org/Names/Names.asp). For an explanation, see [Scientific names and classifications](#), page 2. This automatically generated extension of authoritative classifications is necessary to ensure that the classification includes all species for which the GBIF network offers data. Please notify GBIF (portal@gbif.org) of any problems that have arisen through this automated process.

Step-by-step guide

1. From the home page of the portal, select Explore Species.



2. The portal will display the top level of the classification tree (Kingdoms). For each entry in the tree, two actions are available:

Classification
according to: The Global Biodiversity Information Facility: GBIF Data Portal Classification (based on Catalogue of Life Annual Checklist from specimen and observation data resources)

Find a scientific name within this classification

Jump to a name

All Kingdoms

Kingdom: Animalia	→ see overview page
Kingdom: Archaea	→ see overview page
Kingdom: Bacteria	→ see overview page
Kingdom: Chromista	→ see overview page
Kingdom: Fungi	→ see overview page
Kingdom: Plantae	→ see overview page
Kingdom: Protozoa	→ see overview page
Kingdom: Viruses	→ see overview page

Clicking on the “overview” link to the right of the name will open the Overview Page for that species or group of organisms, which includes access to any species occurrence records available from the GBIF network.

Clicking on the plus symbol (+) to the left of the name will open the classification view that is subordinate to that group of organisms or species (if any). This allows the classification to be explored directly.

3. The classification view also includes an Actions box with quick links for the currently selected species or group of organisms:

free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

search species/...

HOME SPECIES COUNTRIES DATASETS OCCURRENCES SETTINGS

Classification of **Kingdom: Plantae**
according to: The Global Biodiversity Information Facility: GBIF Data Portal Classification (based on Catalogue of Life Annual Checklist, with provisional from specimen and observation data resources)

Actions for Kingdom: Plantae

View: Information about Plantae

Explore: Occurrences of Plantae Classifications of Plantae

Send: Feedback to The Global Biodiversity Information Facility on the classification of Plantae

Information about opens the Overview Page for the current selection.

Explore Occurrences opens the Occurrence Search page so that the user can access the occurrence records that are available for the selection.

Explore Classifications opens the Taxon Search page so you can compare the classification of the selected species or group of organisms used in different data sets available through the portal.

4. Entering a name in the Search field below the Actions box and clicking "Search" will open the Classification Browser view for the species or group of organisms entered.

Find a scientific name within the classification from: GBIF Data Portal Classification (derived from data)

Jump to a name Search

All Kingdoms

- Kingdom: Plantae → see overview page
- Phylum: Anthocerotophyta → see overview page
- Phylum: Bacillariophyta → see overview page
- Phylum: Bryophyta → see overview page
- Phylum: Chlorophyta → see overview page
- Phylum: Cyanidiophyta → see overview page
- Phylum: Cycadophyta → see overview page
- Phylum: Equisetophyta → see overview page
- Phylum: Ginkgophyta → see overview page
- Phylum: Glaucophyta → see overview page
- Phylum: Gnetophyta → see overview page
- Phylum: Hepatophyta → see overview page
- Phylum: Lycopodiophyta → see overview page

Example: Find Komodo Dragons by browsing the classification

To find records of Komodo Dragons by browsing the taxonomic hierarchy, use the following path:

Kingdom: Animalia
Phylum: Chordata
Class: Reptilia
Order: Squamata
Family: Varanidae
Genus: Varanus
Species: *Varanus komodoensis*

Click "see overview page" to go to the Species Overview Page. From there, it is possible to explore occurrences, list datasets providing occurrence records, and download records or map information.

free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

Classification of **Species: Varanus komodoensis** OUWENS
according to: The Global Biodiversity Information Facility: GBIF Data Portal Classification (based on Catalogue of Life Annual Checklist, with provisional from specimen and observation data resources)

Actions for Species: Varanus komodoensis 1912

View: Information about *Varanus komodoensis*

Explore: Occurrences of *Varanus komodoensis* Classifications of *Varanus komodoensis*

Send: Feedback to The Global Biodiversity Information Facility on the classification of *Varanus komodoensis*

Find a scientific name within the classification from: GBIF Data Portal Classification (derived from data)

Jump to a name Search

All Kingdoms

- Kingdom: Animalia → see overview page
- Phylum: Chordata → see overview page
- Class: Reptilia → see overview page
- Order: Squamata → see overview page
- Family: Varanidae → see overview page
- Genus: *Varanus* → see overview page
- Species: *Varanus komodoensis* → see overview page

Finding information about a species or group of organisms

The species or group Overview page

Searching for a species or browsing the classification leads to the Overview Page for the species or chosen group of organisms.

Links to groupings higher than the one displayed (e.g. Genus, Family, Order, Class, Phylum or Kingdom) open an overview of records for all of the species included in that group.

The Overview Page summarises the data available, and includes links to allow exploration of the data in more detail

Several of the links on Overview pages lead to the Occurrence Search page. See Searching for Occurrences, page 16, for more information on using that page.

Free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

Species: *Passer domesticus* (Linnaeus, 1758)
House Sparrow

» Kingdom: Animalia » Phylum: Chordata » Class: Aves » Order: Passeriformes » Family: Passeridae » Genus: Passer

Actions for *Passer domesticus*

Explore: Occurrences Names and classification
List: Countries with occurrences Datasets with occurrences
Download: Darwin Core records One-degree cell density overlay for Google Earth

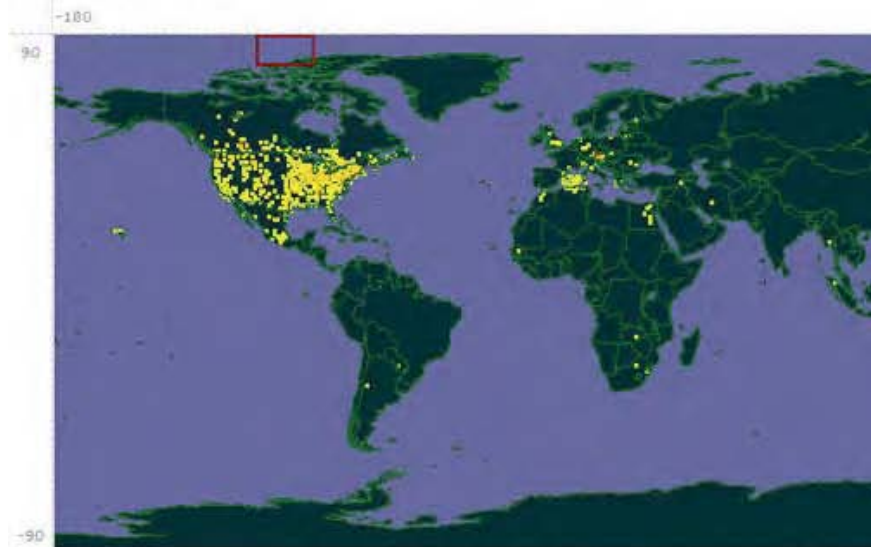
Names and classification

According to Catalogue of Life: 2007 Annual Checklist: The Integrated Taxonomic Information System

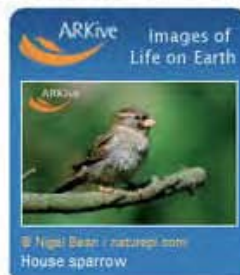
Name	<i>Passer domesticus</i> (Linnaeus, 1758)
Classification	» Kingdom: Animalia » Phylum: Chordata » Class: Aves » Order: Passeriformes » Family: Passeridae » Genus: Passer
Status	Accepted name
Common names	English House Sparrow French Moineau Domestique Spanish Gorri7n Casero
Record identifier	ITS-179628
Record URL	http://www.itis.gov/servlet/SingleRpt/SingleRpt?search_topic=TSN&search_results=179628
Feedback	Feedback to Catalogue of Life: 2007 Annual Checklist on the classification of <i>Passer domesticus</i>

The Names and Classification sections provide information about the organisms and identifies the accepted name.

Occurrence overview



Map only shows records with coordinates (14,712 records from a total of 20,672 records). Map includes data shared for all subspecies included in the species *Passer domesticus* (1 subspecies).



When images or other additional data or information are available, links to these resources are included.



The scientific name of the species or group is given at the top of its Overview Page, along with its place in the overall classification. The names in the classification can be used as quick links to the Overview Pages for those groups of organisms.

Genus: *Passer* » Species: *Passer domesticus*

Placemarks for Google Earth (limit 10,000)

The page includes an **Actions** box with quick links for the current species or group of organisms:

Explore occurrences opens the **Occurrence Search** page, which shows all available occurrence records associated with the species or group.

Information System

Formes » Family: Passeridae » Genus: *Passer* » Species: *Passer domesticus*

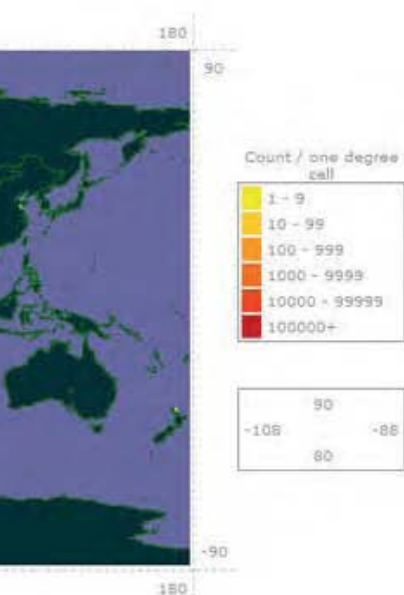
Explore names and classification opens the **Classification Browser** view for the species or group of organisms.

This section provides detail on the names used for the authority for this information where applicable.

There are also options for making lists of countries and datasets and downloading records.

Search_value=179628

Classification of *Passer domesticus* (Linnaeus, 1758)



The **Occurrence overview** section provides a map of all *available* georeferenced records, and links to the occurrence search page for the organisms.

The map includes only *georeferenced* records (those that include map coordinates). There may be additional records available that are not georeferenced. The counts of each type of record are provided below the map.

See [Maps in the GBIF portal](#), page 4, for more about how to use these maps.

(species).

Finding information about a country

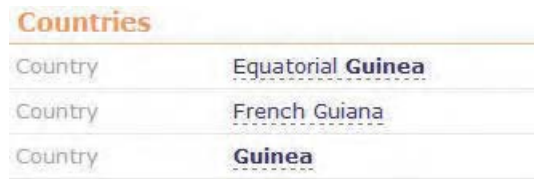
The GBIF Portal provides access to information on the distribution of species in the form of “species occurrence records” (the key details of an occurrence of a particular species at a particular place at a particular time). One of the views that the portal offers is a summary of these data by country of occurrence.

Searching for species that occur within a country

The simplest way to find species data for a country is to enter the name of the country into the search box included on each page.

Step-by-step guide

- From the home page of the portal or any other page, enter the name of a country in the Search box and click **Search**. The search will find the country with that name as well as those countries with names that include the search string. That is, a search on “Guinea” will return species records for Guinea, Guinea-Bissau, Equatorial Guinea, and Papua New Guinea.
- The portal will return lists of names that match the search string. Results are grouped into four categories:
 - Scientific names
 - Common names
 - Countries**
 - Datasets
- Select a name from the **Country** category to see the country’s Overview Page.



Browsing the country list

Alternatively, it is possible to select a country by browsing an alphabetical list. To access the **Country browser**, select the **Explore countries** link on the home page of the portal or the **COUNTRIES** link included in the banner at the top of every page.



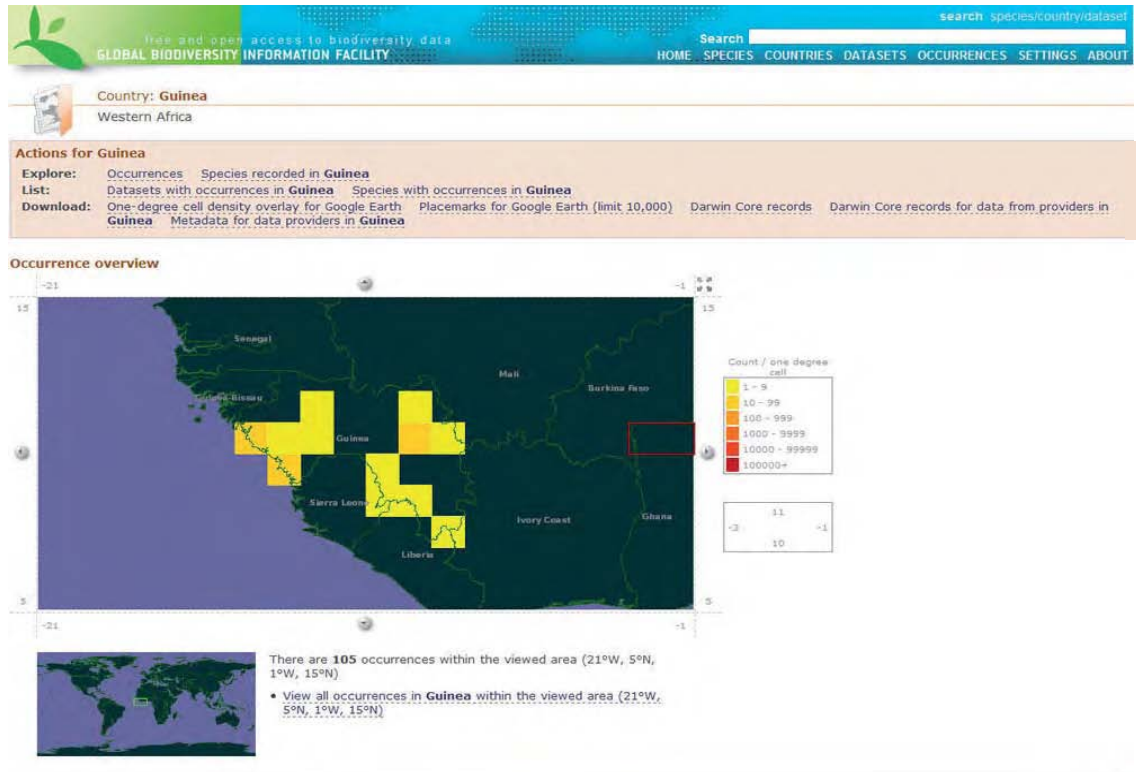
Step-by-step guide

- From the home page of the portal, select **Explore countries**.
- The portal will display a list of countries with names beginning with the letter “A” and links to pages for each other letter of the alphabet. For each country the list gives a count of the number of species occurrence records accessible through the portal, the number of these that have coordinates, and a count of the species known to occur in the country on the basis of these records.
- Select a country name to see its **Country overview** page.



The country Overview page

The Country overview page summarises available information on the species that occur in the country. Several of the links on this page lead to the Occurrence search page - see [Searching for occurrences](#), page 16, for more information on using that page.



Step-by-step guide

1. The Country overview provides a map for all available georeferenced records from the country (all species). Clicking on the map will zoom to higher detail (down to a display showing the density of records for the cell size 0.1 by 0.1 degree); attempting to zoom further then links directly to the Occurrence search page, from whence one can explore the records that occur in the selected cell.
2. The map includes only records with coordinates (remember that there may be other records for the country without coordinates) that occur within the boundaries of the country. Georeferenced records can be seen by selecting the link to "View all occurrences" from the country. See [Maps in the GBIF portal](#), page 4, for more information on using these maps.
3. The page includes an **Actions** box with quick links for the country
 - Explore occurrences** opens the Occurrence search page to see occurrence records identified as occurring in the country. This is a combination of records that include the name of the country as the country of occurrence, as well as other records with coordinates that fall within the country.
 - List datasets with occurrences in ...** opens a list of data providers that are sharing occurrence data relating to the country. It is possible to select one or more providers from this list and view only their records for the selected country. For an example of this, see [Finding information about datasets](#), page 20, that contain records from country of your choice.
4. Download options include downloading records, formatting the download for display in Google Earth, getting more information about the providers of data relating to the selected country, and data providers located in that country.

Example: Ecuador plant species

There are many reasons a person might want to have a list of species of a particular group of organisms that occur in his or her country. In this case, the person is from Ecuador, and he is interested in the plants of his country. Here is how he can get a list, using the Country Overview Page and the Occurrence Search:

Finding plant species of Ecuador



1. Enter "Ecuador" into the search box on the Data Portal.



2. Select Ecuador from the Countries section on the Ecuador search results page to go to the Overview Page for Ecuador.



3. Click on Explore occurrences in the Actions for Ecuador section to open the Occurrence Search for records from Ecuador.

4. In order to limit the search to records of plant species, add a filter for classification. Select Classification from the drop-down menu below Add search filter to open the classification wizard.



5. Click "Kingdom:Plantae" to expand the tree from the selected node. This will create a filter for "Classification includes Kingdom:Plantae."

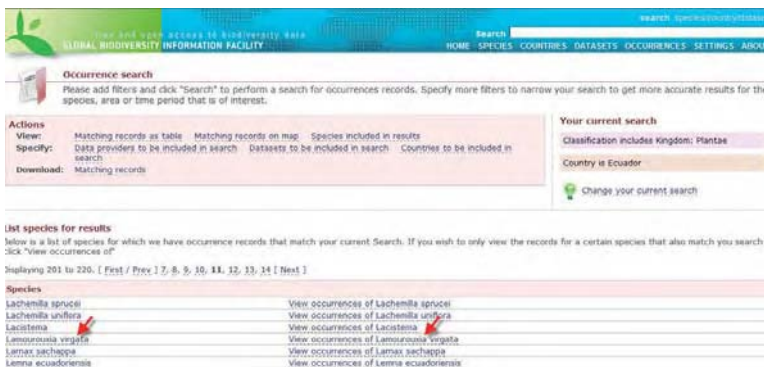
6. Click the Add Filter button to add the filter that is currently displayed. The new filter will appear in the Your current search section.



7. Click the Search button to apply the new filter to the search.



8. Click "Species included in results" in the "Actions" box to retrieve a list of plant species for which occurrence records from Ecuador can be accessed through the GBIF data portal.



9. Clicking on the name of a species in the list will open the Species Overview Page, which provides global information for the species.

10. Click "View occurrences of ..." to view occurrence records for the given species from within Ecuador.

Example: Exploring occurrences recorded within Indonesia

In this example, use the Country Overview Page to reach the Classification Browser for species that occur within Indonesia, and use the Country Overview Page with the Occurrence Search to find datasets from around the world that contain data recorded in Indonesia.

Finding a species within Indonesia



1. Click on the COUNTRIES link from within the Data Portal to browse country names.



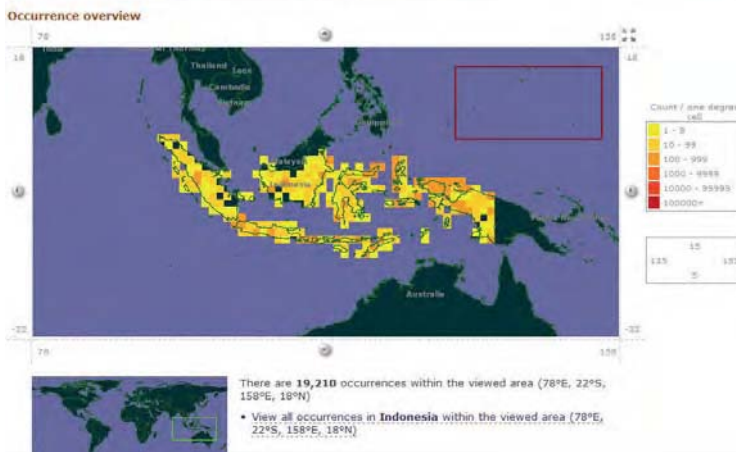
2. Click on the letter I at the top of the Geography Browser page to open the list of country names beginning with I.



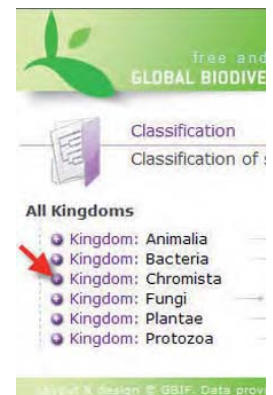
3. Click on "Indonesia" to open the Country Overview Page.



4. Click "Explore species recorded in Indonesia" in the "Actions for Indonesia" section to open the Classification browser restricted to species recorded in Indonesia.



5. Open stages of the classification tree by clicking on the "plus" sign to the left of the name.



Finding dataset(s) with occurrences in Indonesia

1. Click "List datasets with occurrences in Indonesia" in the Actions for Indonesia box to open the corresponding Occurrence search page.
2. Datasets that are providing occurrences recorded in Indonesia are listed on this page, with the associated data provider given below in gray. Click the name of a dataset to open the Dataset Overview Page to find more information about it.
3. To search for occurrences recorded within Indonesia and provided by a particular dataset, tick the check box next to its name (it is also possible to select several datasets at once), and click the Refine search button at the bottom of the page.

Include Dataset	Occurrences
<input type="checkbox"/> Fishes [NHM] The Swedish Museum of Natural History (SMNH)	350
<input type="checkbox"/> Lund Botanical Museum [LD] The Swedish Museum of Natural History (SMNH)	14
<input type="checkbox"/> BCCM/MUCL - (Agricultural) Fungi and Yeasts collection BASF (Private)	32
<input checked="" type="checkbox"/> Herpetology [NHM] The Swedish Museum of Natural History (SMNH)	97
<input type="checkbox"/> Pteridophytes [R] The Swedish Museum of Natural History (SMNH)	1,681

Searching for Occurrences

Access to species occurrence records is at the core of the GBIF Data Portal. The functions provided by the **Occurrence Search** page provide the ability to perform complex searches in order to explore and locate records of interest.

Click on OCCURRENCES at the top of any page in the Data Portal to access the **Occurrence Search** page. Many other pages in the portal provide links to this page and set some initial search filters. For example, the **Explore occurrences** link from a species' **Overview Page** opens the **Occurrence Search** to find records for that species.



Using Search Filters

However the Occurrence Search page is reached, it is always possible to modify a search by adding or removing search filters. Adding filters restricts the occurrence records returned to only those that match the criteria specified by the filter(s). For example, a filter can be applied to limit results to plants in the family Poaceae and then a second filter to further restrict results to records from South Africa. As more filters are added they refine the search and reduce the range of matching records.

As a refined search is being constructed, a list of filters in effect is displayed in the upper right area of the page. A filter can be removed from the list by clicking on the "minus" symbol to the right of the filter.

If a filter category is applied twice (e.g. Scientific name is *Panthera leo* and Scientific name is *Panthera tigris*), the portal will treat them as alternatives and therefore will return records that match either (that is, the search will be for *Panthera leo* OR *Panthera tigris*).

A list of all the filter categories available can be found on the opposite page.

Step-by-step guide

1. Select OCCURRENCES at the top of any page to open the **Occurrence Search** page.
2. Select a filter category (Scientific name, Classification, Country, etc.) from the drop down menu under "Add search filter" (see **Ecuador plants** example, page 16). The portal will automatically provide suitable fields for specifying the value for the filter.
3. Set the filter by clicking "Add filter". This choice is then displayed as part of the search definition under "Your current search" in the upper right of the page.
4. When all desired filters have been set, click on "Search" at the bottom of the list of filters that are in effect.
5. The portal will return an indication of the number of records that match the search (exact numbers for small result sets, otherwise an indication that the count exceeds 1000 records) and will display some sample results (up to 5 records).
6. At this point, you can change or refine the search and issue it again, or begin using the results. The portal displays a set of possible actions:
 - a. **View records on a map** - display the results of the search as a map, which can be explored further by zooming in to different areas. See **Maps in the GBIF portal** for more information on using these maps.
 - b. **View records as a table** - page through the results with the opportunity to open up the **Occurrence Detail** view for any occurrence record (with full details from the record, a map showing the location and links to additional actions, see **Viewing details of an occurrence record**).
 - c. **Download results** - download the search results as XML or as a comma-delimited file.
 - d. **Specify** - Choose to view only those records from a particular country, data provider or dataset.
 - e. **List** - the species resulting from the search.

Available Filter Categories

- **Scientific name** - enter a scientific name (or part of a name) and choose between "is" and "is like". This filter will return any records that have a matching name given for the identification of the organism, regardless of how the organism is classified.
- **Common name** - return any records for species that are associated with the common name supplied (if that common name and its scientific equivalent has been provided to GBIF).
- **Classification** - select a species or group of organisms from the classification tree. This filter will return records for any species within the selected part of the tree.
- **Type status** - find specimens that are marked as types.
- **Country** - select a country from the pick list. This filter will return records from the country identified, whether or not they contain coordinates (but note that adding a coordinate filter (Bounding box, Latitude or Longitude) will limit the results to georeferenced records).
- **Region** - select a geographic region from the pick list. This filter will return records from all countries in the selected continent, oceanic area, etc.
- **Bounding box** - use the map to select a rectangle defined using latitudes and longitudes. This filter will return only georeferenced records within the given rectangle.
- **Latitude** - enter a latitude and choose between "is", "greater than" and "less than". This filter will return only georeferenced records that match the selection.
- **Longitude** - enter a longitude and choose between "is", "greater than" and "less than". This filter will return only georeferenced records that match the selection.
- **Coordinate status** - select "includes coordinates" to filter out those records that are not georeferenced; alternatively, select "does not include coordinates" to exclude georeferenced records. To see all records, do not use this filter.
- **Coordinate issues** - select "issues detected" to identify records with coordinates that may be doubtful; otherwise, select "issues not detected".
- **Data provider** - select a data provider from the pick list. This filter will return records from the specified provider.
- **Host country** - select a country from the pick list. This filter will return records that are being shared by data providers in that country.
- **Dataset** - select a data network or a data provider from the first pick list, and then a dataset from the second pick list. This filter will return records from the selected dataset.
- **Occurrence date** - enter a start date and either enter an end date or select "use specific date". This filter will return records occurring on the specified date or in the specified date range.
- **Year Range** - select records from a range of years.
- **Year** - select a year and either "is", "before" or "after". The filter will return records from the year(s) that match the selection.
- **Month** - select a month from the pick list. The filter will return records from the specified month, regardless of year.
- **Institution code** - enter an institution code and "is" or "is like". The filter will return records with the specified institution code.
- **Collection code** - enter an collection code and "is" or "is like". The filter will return records with the specified collection code.
- **Catalogue number** - enter an catalogue number and "is" or "is like". The filter will return records with the specified catalogue number.
- **Basis of record** - select a basis of record (specimen, observation, living, germplasm, fossil) from the pick list. The filter will return records that are based on the selected type of object or observation.

Viewing details of an Occurrence record

Occurrence records can be found using the Occurrence search (see Searching for occurrences, page 16). In the tabular view offered by the Occurrence search, there is a View link shown to the right of each record. This link opens the Occurrence detail view, which in turn offers a link to retrieve the original record directly from the provider's web site.

Step-by-step guide

1. Perform a search using the Occurrence search (page 16) and then select the option to View matching records as table.

Table of results

Displaying 261 to 280. [First / Prev] 10, 11, 12, 13, 14, 15, 16, 17 [Next]

Dataset	Scientific Name	Institution Code	Collection Code	Catalogue No.	Coordinates	Date	Country	
Herts Bird Club - Hertfordshire Breeding...	<i>Turdus merula</i>	Herts Bird Club	3372	25058113	51.687984, -0.410119		United Kingdom	View
Herts Bird Club - Hertfordshire Breeding...	<i>Turdus merula</i>	Herts Bird Club	3372	25058155	51.777863, -0.406962		United Kingdom	View
Herts Bird Club - Hertfordshire Breeding...	<i>Turdus merula</i>	Herts Bird Club	3372	25058142	51.777863, -0.406962		United Kingdom	View
Herts Bird Club - Hertfordshire Breeding...	<i>Turdus merula</i>	Herts Bird Club	3372	25058157	51.777863, -0.406962		United Kingdom	View
Herts Bird Club - Hertfordshire Breeding...	<i>Turdus merula</i>	Herts Bird Club	3372	25058168	51.867737, -0.403789		United Kingdom	View
Herts Bird Club - Hertfordshire Breeding...	<i>Turdus merula</i>	Herts Bird Club	3372	25058178	51.867737, -0.403789		United Kingdom	View

2. Click on the "View" link at the right of a table row to see the Occurrence detail view for that record. A new page will open, showing
 - a. many of the original values supplied by the data provider,
 - b. an interpretation of some of these fields to make it possible to search for records across all datasets, and
 - c. notes on apparent discrepancies within individual records (e.g. inconsistent country and coordinates).

Actions
Find:
Retrieve:
View:
Send:
Dataset
Data Provid
Dataset:
Institution
Collection
Catalogue
Basis of rec
Field numb
Collector n
Collector n
Date collec
Taxonomy
Scientific n
Geospatial
Continent:
Country:
County:
State/Prov
Locality:
Coordinate
Altitude:
Depth:

- There may be a “Warnings” box at the top of the page. This will note any apparent discrepancies within the record, and may highlight that some important data elements, such as basis of record, are not available or could not be interpreted automatically by the portal.
- The information for the record is displayed under a series of subheadings:



Occurrence Details: Herts Bird Club 3372 25058157

Turdus merula

Records of *Turdus merula* from Herts Bird Club - Hertfordshire Breeding Bird Atlas : 1988-1992 [All records of *Turdus merula* from same one degree cell \(0.0°W, 51.0°N, 1.0°W, 52.0°N\)](#) [All records of any species at 51.777863, -0.406962](#)

[Retrieve original record from data provider](#) [Download Google Earth \(KML\) file for this occurrence](#)

Classification of *Turdus merula* according to Herts Bird Club - Hertfordshire Breeding Bird Atlas : 1988-1992

[Feedback to UK National Biodiversity Network](#)

Provider:	UK National Biodiversity Network Herts Bird Club - Hertfordshire Breeding Bird Atlas
Code:	Herts Bird Club
Record code:	3372
Record ID:	25058157
Record type:	O (interpreted as Observation)
Year:	HBBA_1988_
Number:	0
Name:	
Year:	Year: 1988
Name:	<i>Turdus merula</i> Linnaeus, 1758 (interpreted as <i>Turdus merula</i>)
Region:	Europe (interpreted as Europe)
Country:	United Kingdom
Coordinates:	51.777863000000004, -0.4069619999999999

Dataset - metadata on the original record and any specimen it represents.

This section contains all identifiers associated with the record and provides links back to the pages for the data provider and dataset. It also displays the “Basis of record”. This indicates what the data provider is using to show that the organism did occur at the time and place indicated (e.g. the basis of record is a specimen held in a collection, or an observation or fossil). The portal shows both the original information supplied by the data provider as the basis for this record (e.g. “S”) and how the portal has interpreted that value (e.g. “interpreted as Specimen”). If the portal has not been able to interpret the value, or no value is supplied, the basis of record appears as Unknown.

Taxonomy - information on the identification associated with the record,

including the classification used by the GBIF portal, and any other classification information supplied by the data provider. In each case the portal displays the original information from the provider and offers a link to the overview pages that show the species or group of organisms with which the portal has associated the record - this is particularly significant in cases in which there are multiple organisms with the same name or the classification supplied by the data provider differs significantly from that used by the portal. This section also includes information on the person who identified the organism (if any was included by the data provider).

Geospatial - information on the location and date of the occurrence. The

portal again displays both the information retrieved from the data provider and the interpretation of that information by the portal. If the record has coordinates, a Google Map view is also shown for the location. This section also includes any information on the collector or observer (if this was included by the data provider).

- In the Actions box, there are several links, including one that retrieves the original record from the data provider (this may contain additional information, or may possibly have been modified since it was indexed by the portal) and options to find related records.

Finding information about a dataset

The GBIF Portal provides access to information from a wide range of different organisations and institutions. One of the views that the portal offers is a summary of the data shared by each data provider or included in an individual dataset or one of the (cross-institution) information networks to which many of these datasets belong.

GBIF data providers are of different types:

1. Providers of information on the names and classification of organisms.
2. Providers of information on the occurrence of organisms in different locations and at different times.
3. Providers of images and other information on these organisms.

The Datasets area of the GBIF portal provides information on all three types of providers and the datasets they are sharing.

Searching for a dataset



The simplest way to find information on a dataset, data provider, or data network is to search for it by name using the Search box included on each page.

Step-by-step guide

1. From the home page of the portal or any other page, enter the name of the dataset in the Search box and click Search.
2. The portal will return lists of names that match or include the search string. Results are grouped into four categories:
 - a. Scientific names
 - b. Common names
 - c. Countries
 - d. Datasets
3. Simply select a name from the Dataset category to see the Dataset overview page for the given dataset, data provider or data network.

free and open access to biodiversity data
GLOBAL BIODIVERSITY INFORMATION FACILITY

Search Results for: **Asteraceae**

[Scientific names](#) [Common names](#) [Countries](#) [Datasets](#)

Scientific names

Family	Asteraceae
Genus	Asteraceae
Subfamily	Asteraceae subfam. Barnadesioideae
Subfamily	Asteraceae subfam. Cichorioideae
Subfamily	Asteraceae subfam. Corymbioideae
Subfamily	Asteraceae subfam. Gochnatioideae
Subfamily	Asteraceae subfam. Gymnarrhenioideae
Subfamily	Asteraceae subfam. Hecastocleioideae
Subfamily	Asteraceae subfam. Lactucoideae
Subfamily	Asteraceae subfam. Pertyoideae

[View all scientific names matching "Asteraceae"](#)

Common names

No common names matching "Asteraceae"

Countries

No countries with names matching "Asteraceae"

Datasets

Dataset	Asteraceae
---------	-------------------

Browsing the dataset list

Alternatively, it is possible to select a dataset, data provider, or data network by browsing an alphabetical list. To access the Dataset browser, select the **Explore datasets** link on the home page of the portal or the DATASETS link included in the banner at the top of every page.



Step-by-step guide



1. From the home page of the portal, select Explore datasets.



- The portal will display a list of datasets, data providers, and data networks with names beginning with the letter "A" and links to pages for each other letter of the alphabet. The information is presented in three separate categories (some categories may not be present for all letters of the alphabet):

Data networks - (cross-institution) information networks including datasets from a number of data providers.

Data providers - institutions and organisations serving data through the GBIF network.

Datasets - individual data sets shared by a data provider.

The screenshot shows the GBIF Datasets page. At the top, there is a navigation bar with the text "free and open access to biodiversity data" and "GLOBAL BIODIVERSITY INFORMATION FACILITY". Below this is a "Datasets" section with a navigation bar containing letters A through Z, with 'O' selected. Under the 'O' tab, there are three categories: "Data Networks" (Ocean Biogeographic Information System), "Data Providers" (Ohio State University Insect Collection, Oregon State University, Österreichische Mykologische Gesellschaft, OZCAM (Online Zoological Collections of Australian Museums) Provider, Univ. of Kansas Biodiversity Research Center), and "Datasets" (Observational database of Icelandic plants, Icelandic Institute of Natural History).

The Dataset Overview page

The Dataset Overview page provides information about the institution(s), organisation(s) and dataset(s) involved and summarises the occurrence data in a dataset or in all datasets served by a given data provider or included in a data network. Several of the links on this page lead to the Occurrence search page (see Searching for occurrences, page 16) for more information on using this page.

Step-by-step guide

- Each Dataset overview page provides a map for all available georeferenced records from the dataset, data provider or data network (all species). Clicking on the map will zoom to higher detail (down to a display showing the density of records for each 0.1 by 0.1 degree cell) and then links directly to the Occurrence search page to explore the records from the selected cell.
- The map includes only records with coordinates. Click the "View all" link below the map to open the occurrence search view to see georeferenced records from this dataset that occur within the area currently shown on the map. See Maps in the GBIF portal (page 4) for more information on using these maps.
- The page includes an Actions box with quick links for the dataset, data provider or data network. **Explore occurrences** opens the Occurrence search page to see occurrence records from the dataset, data provider or data network. **Explore names and classification** (only on overview pages for datasets) opens a view of the species and higher groups served by the dataset. Options to list species or countries with occurrences included in the dataset are also included.
- Each Dataset Overview page presents further information on the institutions and organisations involved. Provider and network pages provide links to the relevant datasets. Dataset pages link back to the information on the data provider and to information on any networks to which the dataset belongs.

The screenshot shows the Dataset Overview page for "Fishes in the Argentine Sea from 1967 to the present time". It includes a map of the Argentine Sea with a density heatmap of records. A legend on the right indicates density ranges from 0 to 1000. Below the map, there is a "Metadata" section with the following information:

- Dataset:** Fishes in the Argentine Sea from 1967 to the present time
- Provider:** Institute of Marine and Coastal Sciences, Rutgers University
- Network:** Ocean Biogeographic Information System
- Number of records:** 3,449
- Number of records with coordinates:** 1,121
- Number of species:** 136
- Number of taxa:** 378

There is also a "Metadata" section with the following information:

- Dataset:** Fishes in the Argentine Sea from 1967 to the present time
- Provider:** Institute of Marine and Coastal Sciences, Rutgers University
- Network:** Ocean Biogeographic Information System
- Number of records:** 3,449
- Number of records with coordinates:** 1,121
- Number of species:** 136
- Number of taxa:** 378

Example: Finding datasets that have georeferenced records

To answer a certain question, an environmental planner needs a large number of data records that can be mapped. He can find datasets that meet his requirements by doing the following.

1. Open the Occurrence Search Page by clicking on OCCURRENCES in the banner on any page of the portal.
2. In the Occurrence Search page, under "Add search filter", choose "Coordinate Status" in the drop-down menu, and that the right-hand drop-down is on "Includes coordinates". Click "Add Filter".
3. To the right, under "Your current search", click Search. This will limit the results list to only those records that can be mapped, because they contain geographic coordinates. Some of these are listed in a "Sample results" table on the page that appears.
4. Now, choose to "Specify Datasets to include", by clicking on the link in the Actions box.

Occurrence search
Please add filters and click "Search" to perform a search for occurrences records. Specify more filters to narrow your search for species, area or time period that is of interest.

Add search filter

Coordinate status Includes coordinates

Your current search

Scientific name is Chiroptera

Coordinate status is Includes coordinates

This search matches 324 occurrence records.

Actions

View: [Matching records as table](#) [Matching records on map](#) [Species included in results](#)

Specify: [Data providers to be included in search](#) [Datasets to be included in search](#) [Countries to be included in search](#)

Download: [Matching records](#)

Sample results

Dataset	Scientific Name	Institution Code	Collection Code	Catalogue No.
The Swedish Museum of Natural History...	<i>Chiroptera</i>	NRM	Mammals	607006
The Swedish Museum of Natural History...	<i>Chiroptera</i>	NRM	Mammals	611345
The Swedish Museum of Natural History...	<i>Chiroptera</i>	NRM	Mammals	642262
The Swedish Museum of Natural History...	<i>Chiroptera</i>	NRM	Mammals	642263
The Swedish Museum of Natural History...	<i>Chiroptera</i>	NRM	Mammals	875239

5. On the resulting page, click on "Specify Datasets to be included" in the Actions box. In the page that opens, click on the name of a dataset or a data provider to go to its Overview Page. Or, to see only those records from one or more datasets, choose them by ticking the box to the left, and then on "Refine search".

Specify the datasets to include in your search
Below is a list of datasets that have provided occurrence records that match your current search. If you wish to only view the records from a checkbox and click "Refine search".

Include Dataset

Mammals (NRM)
The Swedish Museum of Natural History (NRM)

Paleobiology Database
Marine Science Institute, UCSB

Natural England - Batsites inventory for Britain
UK National Biodiversity Network

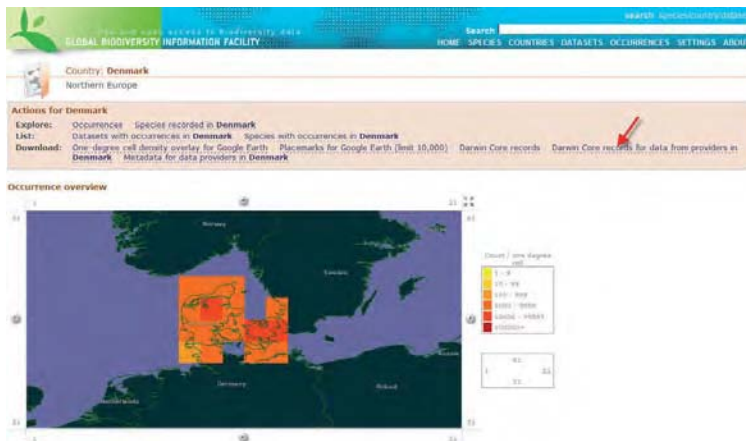
Example: Finding data providers from Denmark

A Danish biology student needs to know what institutions in her country are sharing biodiversity data via the GBIF network. Here is how she would find out.

1. Click on the COUNTRIES link in the Data Portal.
2. Click on the letter D at the top of the Geography Browser page to open the list of country names beginning with D (countries are listed alphabetically by their English names).



3. Click on "Denmark" to open the Country Overview page.
4. In the "Actions" box, click on "Download metadata for data providers in Denmark" to retrieve a list of data providers from Denmark.



5. The request summary at the top of the Provider Web Service Response page shows that records for two data providers from Denmark have been returned.

GBIF occurrence web service response

Request details	
service	occurrence
hostisocountrycode	DK
maxresults	500
mode	processed
request	list
format	darwin
First record returned	0
Next record available	500
Number of records returned	500
Number of records matched	501
Next page of records	http://newportal.gbif.org/ws/rest/occurrence/list?hostiso

Copyright Global Biodiversity Information Facility, 2007
This document contains data from the GBIF Network - see <http://www.gbif.net/> for more information.
All usage of these data must be in accordance with the GBIF Data Use Agreement - see <http://www.gbif.org/>

For help with this web service, see: <http://newportal.gbif.org/ws/rest/occurrence/help>

Occurrence data

Data provider: Herbarium of the University of Aarhus

Data provider key in GBIF portal: 69
Data provider page in GBIF portal: <http://newportal.gbif.org/datasets/provider/69>
Web service request for data provider: <http://newportal.gbif.org/ws/rest/provider/69>

Dataset: The AAU Herbarium Database

rights: None
Citation: AAU Herbarium Database
Dataset key in GBIF portal: 224
Dataset page in GBIF portal: <http://newportal.gbif.org/datasets/dataset/224>

6. In the Metadata for GBIF data providers section of the Provider Web Service Response page, data providers are listed first, and their datasets are listed next. A brief description is given for each provider, along with a URL for the data provider's Web site and numeric counts of occurrence records and taxonomic units. Clicking the link given at *Data provider page in GBIF portal* will open the Data Provider Overview Page.
7. Each dataset that is made available by a data provider is listed on the Provider Web Service Response page. A brief description is given for the dataset, along with numeric counts of records and information on access points. Click the link given at *Dataset page in GBIF portal* to open the Dataset Overview Page.
8. Associated taxon concepts are listed below the corresponding dataset. The name, rank, and source of the taxon concept are given. Click on the link given at *Taxon page in GBIF portal* to open the Taxon Overview Page.

Example: Migratory birds

A person interested in birds might wish to know where the swallow that nests in his barn in spring and summer goes during the winter. This example shows how to map records made at different times of the year to answer the question.

1. Enter "barn swallow" into the search box on the Data Portal.
2. Click on "Species: *Hirundo rustica* (English: Barn Swallow)" in the Common names section of the search results page to open the species' Overview Page.



Species: ***Hirundo rustica* Linnaeus, 1758**
Barn Swallow

Kingdom: [Animalia](#) » Phylum: [Chordata](#) » Class: [Aves](#) » Order: [Passeriformes](#) » Family: [Hirundinidae](#)

Actions for *Hirundo rustica*

Explore: [Occurrences](#) [Names and classification](#)

List: [Countries with occurrences](#) [Datasets with occurrences](#)

Download: [Darwin Core records](#) [One-degree cell density overlay](#)

3. Click **Explore occurrences** in the **Actions for *Hirundo rustica*** section to open the Occurrence Search page for *Hirundo rustica*.

4. In order to limit the search to occurrences recorded during a particular season, build a filter for a month (see Searching for Occurrences, page 16). Select "Month" from the drop-down menu below "Add search filter". Select "January" from the drop-down menu that provides month options for the filter, and click the "Add filter" button.
5. Add another filter for "Coordinate status" is "Includes coordinates".
6. Add another filter for "Coordinate issues" is "No issues detected".

Your current search

Classification includes Species: *Hirundo rustica*

Coordinate status is Includes coordinates

Coordinate issues is No issues detected

Month is January

Search

7. Click the **Search** button to retrieve the filtered results.
8. Click **View matching records on map** in the **Actions** section to open a page with a map displaying occurrences of barn swallows recorded in January.

This search matches 343 occurrence records.

Actions

View: [Matching records as table](#) [Matching records on map](#)

Specify: [Data providers to be included in search](#) [Datasets to include](#)

Download: [Matching records](#)


Sample results

Dataset	Scientific Name	Institution
Biologiezentrum der Oberoesterreichische...	<i>Hirundo rustica</i>	LI
Biologiezentrum der Oberoesterreichische...	<i>Hirundo rustica</i>	LI



Your current search

- Classification includes Species: Hirundo rustica
- Coordinate status is Includes coordinates
- Coordinate issues is No issues detected
- Month is January

 [Change your current search](#)

- Right-click on “Change your current search” to open the Occurrence Search page for Barn Swallows in a new browser window or tab. Keep the January map open in the original tab or window.

- In the new instance of the Occurrence Search page, click the “minus” sign next to “Month is January” to remove this filter from the search.
- Add a new “Month” filter for “July” and click the “Search” button to retrieve new results.

- Click “View records on map” in the “Actions” box to open a page with a map for July occurrences of barn swallows.

This search matches over 1,000 occurrence records.

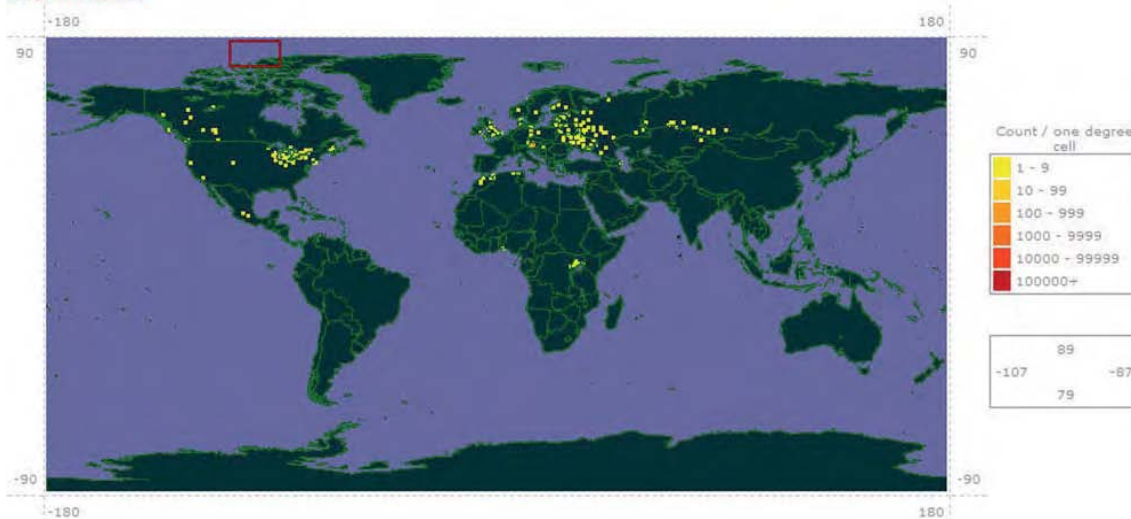
Actions

View: [Matching records as table](#) [Matching records on map](#)

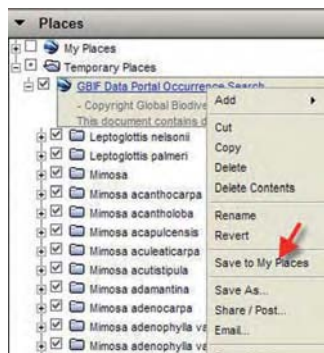
Specify: [Data providers to be included in search](#) [Datasets to be included](#)

Download: [Matching records](#)

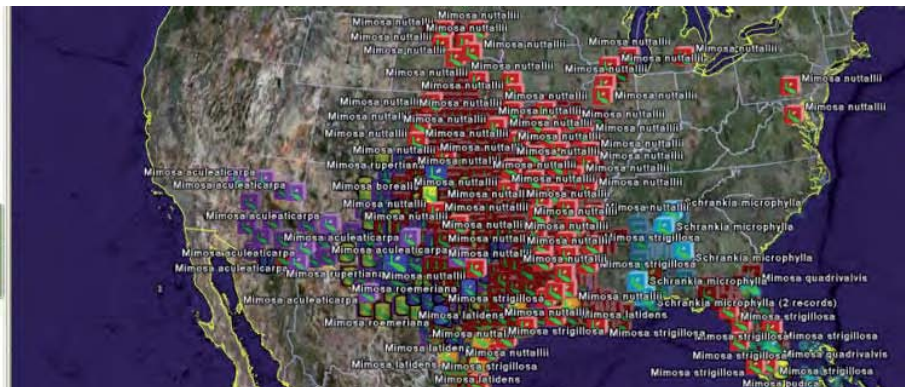
Map of results



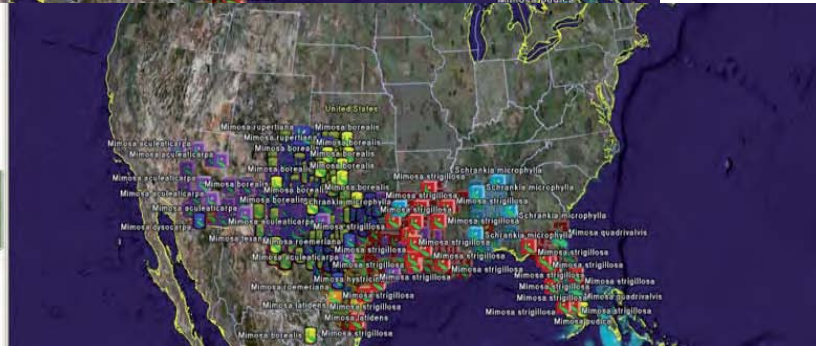
For more information on using the Google Earth interface, consult the Google Earth User Guide (earth.google.com/userguide/v4/).



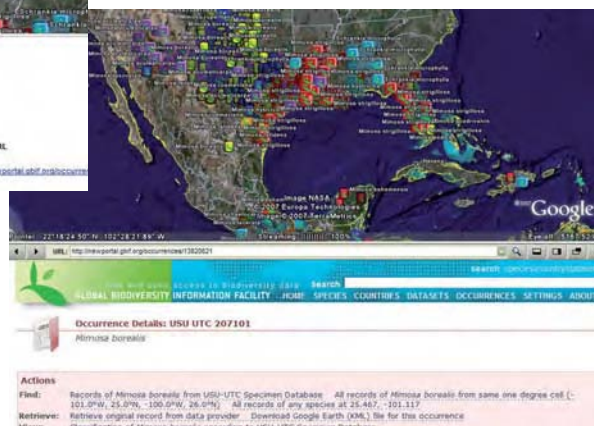
6. Saving placemarks:
 - a. To save the placemarks for the entire genus, right-click on "GBIF Data Portal Occurrence Search" in the Places pane and select "Save to My Places".
 - b. To save the place marks for an individual species, right-click on the species name and select "Save to My Places".
7. The check-boxes in the Places pane of the left-hand sidebar of the Google Earth interface can be used to turn on and off the display of individual species. Scroll down in the pane to find *Mimosa nuttallii*. Click the check box to the left of the name to turn the display of the species off. Choose other species to turn off (or on).



8. Click on any individual icon on the map.
 - a. If the icon represents a single data record, a dialog box will open that gives information on the data provider, and a "portal URL" for the record represented by the icon.
 - b. If the icon represents multiple data records, a branching diagram will appear with single-record icons at the tips of the branches. Click on one of these.



9. Click on the "portal URL" in the dialog box, and an Internet browser pane will open at the bottom of the Google Earth window that shows the GBIF portal Occurrence Search Detail View page for the record selected.



Hummingbird record density

Scenario: An ecologist who is searching for indicator species to use in a biodiversity status assessment needs to find out if there is enough data available via GBIF to use hummingbirds as indicator species. This walk-through demonstrates how she can assess occurrence record density (that is, find out how many data records there are) for hummingbirds in North America by plotting results from the GBIF portal on Google Earth.

1. Enter "Trochilidae" into the search box on any page of the portal.
2. From the search results page, click on "Family: Trochilidae" to open the Overview Page for the bird family Trochilidae (hummingbirds).
3. In the Actions for Trochilidae box, click on "Download 1-degree cell density overlay for Google Earth".
4. Switch to Google Earth, and open the downloaded file.
5. Rotate the view in Google Earth to see the overlay of occurrences in North America.
6. Adjust the transparency of the overlay with the slider located below the Places section of the left sidebar. Note that the density of records in each cell is indicated by colour-code, as it is on maps that are generated within the GBIF portal. For explanation, see Maps in the GBIF portal, page 4.
7. To save the layer, right-click on the description in the Places pane within the left sidebar and select "Save to My Places".

The image shows a composite of two screenshots. The left screenshot is from the GBIF portal, displaying a search result for 'Trochilidae'. The 'Places' sidebar on the left lists various data layers with record counts. The 'Actions for Trochilidae' section includes options like 'Occurrences', 'Names and classification', 'List', 'Download', 'Darwin Core records', 'One-degree cell density overlay for Google Earth', and 'Placemarks for Google Earth'. A red arrow points to the 'One-degree cell density overlay for Google Earth' link. Below this, the 'Names and classification' section shows the taxonomic hierarchy: Kingdom: Animalia, Phylum: Chordata, Class: Aves, Order: Apodiformes, Family: Trochilidae. The 'Occurrence overview' section shows a world map with yellow dots representing records, primarily concentrated in North and South America.

The right screenshot is from Google Earth, showing a 3D view of the Earth with a yellow and green density overlay on the continents of North and South America. The overlay indicates the density of hummingbird records in 1-degree cells. The Google Earth interface includes a navigation pad, a scale bar, and copyright information: 'Image NASA © 2007 Europa Technologies Image © 2007 TerraMetrics'. The Google logo is visible in the bottom right corner.

Booklet and CD © 2007 Global Biodiversity Information Facility (GBIF).

Permission to copy and/or distribute all or part of the information contained herein is granted, provided that such copies carry due attribution to the Global Biodiversity Information Facility (GBIF).

Printing history

First printing June 2007

Second printing (with additions and corrections) September 2007



**GLOBAL
BIODIVERSITY
INFORMATION
FACILITY**

SPECIES COUNTRIES DATASETS OCCURRENCES SETTINGS ABOUT

```
<?xml version="1.0" encoding="UTF-8
<response xmlns="http://rs.tdwg.org/t
thead>
<source accesspoint="http://145.18.162
<software name="TapirLink" version="0.2(re
```



... free and open access to biodiversity data



**GLOBAL
BIODIVERSITY
INFORMATION
FACILITY**

Secretariat
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark
Tel: +45 35 32 14 70
Fax: +45 35 32 14 80
Email: info@gbif.org
Web: www.gbif.org



The pages that follow contain the documents to which the links in the Introduction of this Manual refer.

These documents are not part of the Manual, but are placed here for the convenience of those readers who are not online.

MEMORANDUM OF UNDERSTANDING
FOR THE GLOBAL
BIODIVERSITY INFORMATION
FACILITY

Approved at GB12 in Cape Town, South Africa, April 2006
Annex 1 approved at GB12.5 in Madrid, Spain, June 2006

Table of Contents

	Page
Preamble	1
Paragraph 1: Definitions	2
Paragraph 2: Understandings	3
Paragraph 3: Objectives	4
Paragraph 4: The Governing Board	6
Paragraph 5: The GBIF Secretariat Host	8
Paragraph 6: The GBIF Secretariat	9
Paragraph 7: The Executive Secretary	10
Paragraph 8: Intellectual Property	11
Paragraph 9: Finance	13
Paragraph 10: Association and Disassociation of Participants	14
Paragraph 11: Other matters	15
Annex 1: Financial Contributions for Voting Participants	16

MEMORANDUM OF UNDERSTANDING
FOR THE GLOBAL
BIODIVERSITY INFORMATION FACILITY

The signers of this non-binding Memorandum of Understanding (MOU), being countries, economies, or inter-governmental or international organisations, or entities designated by them, have decided that a co-ordinated international scientific effort is needed to enable users throughout the world to openly share and put to use vast quantities of global biodiversity data, thereby advancing scientific research in many disciplines, promoting technological and sustainable development, facilitating the conservation of biodiversity and the equitable sharing of its benefits, and enhancing the quality of life of members of society. The importance of making biodiversity data openly available to all countries and individuals is underscored by various international agreements.

Recognising this need, the delegates to the Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level in Paris on 22–23 June 1999 endorsed the recommendation that a Global Biodiversity Information Facility (designated hereafter as GBIF) be established, with open-ended participation.

Noting that GBIF was established in March 2001, and that the first MOU for GBIF had a duration of five years (2001-2006), the signers of this Memorandum of Understanding hereby express their intention either to continue their existing Participation in GBIF or to become new Participants of GBIF as a form of technical and scientific international co-operation.

Paragraph 1
DEFINITIONS

1. Biodiversity Data

In the context of this MOU, biodiversity data refers to scientific data, primarily about biological species and about specimens or observations of individual organisms.

2. Participant

A country, economy, inter-governmental or international organisation or an entity designated by them, that has signed this MOU and has expressed its intention to observe the provisions herein. A Participant may designate an entity to take part in the operation of GBIF and to act for the Participant in such matters as the Participant chooses to delegate to it.

3. Participant Node

A mechanism by which a Participant coordinates and supports its GBIF data-sharing activities. A Participant Node includes both physical infrastructure and human resources. Typically a Participant Node encourages and supports the activities of the Participant's data providers to both contribute and use GBIF-served data, provides information technology (IT) infrastructure and expertise for GBIF activities, and functions as an information gateway among Participants, other partners, and the Secretariat.

Paragraph 2
UNDERSTANDINGS

1. GBIF is an open-ended international co-ordinating body set up with the overall aim of furthering technical and scientific efforts to develop and maintain a global information facility for sharing of digital biodiversity data.
2. The Participants' involvement in this MOU is subject to the goodwill and effective contribution to GBIF's activities and operations, either financially or through any of the activities described in Article 3.3, and is subject to the applicable laws and regulations of the Participants.
3. Nothing in this MOU should be read to contradict the principles of the Convention on Biological Diversity and other relevant Conventions.
4. This MOU continues the goals and intents of the original GBIF MOU (2001-2006), in order to sustain the benefits of GBIF and allow its continued existence and development. The Governing Board may decide on suitable arrangements to facilitate the continued involvement and transition of the Participants from the original MOU to this new one.
5. This document is not legally binding and will have no effect as a legal or political precedent.

Paragraph 3

OBJECTIVES

1. Purpose

The purpose of GBIF is to promote, co-ordinate, design and implement the compilation, linking, standardisation, digitisation and global dissemination of the world's biodiversity data, within an appropriate framework for property rights and due attribution. GBIF will work in close co-operation with established programmes and organisations that compile, maintain and use biological information resources. The Participants, working through GBIF, will establish and support a distributed information system that will enable users to access and utilise considerable quantities of existing and new biodiversity data.

2. Goals of GBIF

It is the intention of the Participants that GBIF:

- (a) be shared and distributed, while encouraging co-operation and coherence;
- (b) be global in scale, though implemented nationally and regionally;
- (c) be accessible by individuals anywhere in the world, offering potential benefits to all, while being funded primarily by those that have the greatest financial capabilities;
- (d) promote standards and software tools designed to facilitate their adaptation into multiple languages, character sets and computer encodings;
- (e) serve to disseminate technological capacity by drawing on and making widely available scientific and technical information; and
- (f) make biodiversity data universally available, while fully acknowledging the contribution made by those gathering and furnishing these data.

3. Involvement of the Participants

Each Participant signing this MOU should seek to:

- (a) participate actively in the formulation and implementation of the GBIF Strategic Plan and the Work Programme;
- (b) share biodiversity data through GBIF under a common set of technical standards and within an Intellectual Property Rights framework (such as that described in Paragraph 8);
- (c) form a Participant Node or Nodes, accessible via GBIF, that will organise and/or provide access to biodiversity data, or to data and metadata standards, software tools or other services to enhance the GBIF network;
- (d) as appropriate, make other investments in biodiversity information infrastructure in support of GBIF, as well as helping to co-ordinate and harmonise the biodiversity informatics programs of the Participants, and
- (e) as appropriate, contribute to training and capacity development for promoting global access to biodiversity data, including implementing specific programs to enhance the biodiversity informatics capacity and technical skills base of developing countries.

4. Co-operation and Co-ordination

The Participants intend to encourage co-operation amongst themselves in the implementation of GBIF and in the development of joint work programmes in areas of mutual interest with the Secretariat of the Convention on Biological Diversity and other appropriate bodies and initiatives to avoid duplication and to benefit from existing resources and expertise.

Paragraph 4
THE GOVERNING BOARD

1. Role and Purpose

The Governing Board will be the means by which the Participants will make collective decisions on all matters relating to GBIF, which will then be put into effect by the GBIF Secretariat.

2. Composition

The Governing Board will consist of one representative from each Participant.

There are two modes of participation:

(a) Voting Participants

Voting Participants are those Participant countries that have decided to make the financial contribution suggested in Annex I and have made a financial arrangement as described in Paragraph 10.2.

(b) Associate Participants

There are two categories of Associate Participants

- (i) Associate Participant Countries: those Participant countries that have not yet decided to make the financial contribution suggested in Annex I. They are eligible and encouraged to become Voting Participants as soon as possible.
- (ii) Associate Participant Organisations and Economies: all Participant economies, intergovernmental organisations, and international organisations are Associate Participants. They are not eligible to become Voting Participants.

Associate Participants are encouraged to take part in the deliberations of the Governing Board, but may not vote.

3. Additional Participants

The Secretariat of the Convention on Biological Diversity is invited to designate a non-voting representative to the Governing Board.

4. Voting

- (a) The Governing Board should work by consensus where mandated in this MOU.
- (b) In all other decisions, the Governing Board should strive to work by consensus whenever possible. If consensus is not reached after reasonable attempts have been made, then decisions will be made by super-majority, unless the Governing Board has decided in its Rules of Procedure to approve by simple majority.
- (c) A super-majority vote is the affirmative vote of a two-thirds majority of the Participants present and voting.
- (d) A simple-majority vote is the affirmative vote of more than one-half of the Participants present and voting.

- (e) In all cases in which this MOU expressly provides that the Governing Board act by means of a consensus decision or a vote of the Participants present and voting, “present” can mean face-to-face, by telephone, Internet, video conference, or other practical means determined in advance by the Governing Board.

5. Responsibilities

The Governing Board may:

- (a) establish its Rules of Procedure and such subsidiary bodies as it sees necessary for its proper functioning and the achievement of GBIF goals;
- (b) form relationships with organisations that may assist GBIF to achieve its goals;
- (c) adopt a multi-annual Strategic Plan for GBIF;
- (d) adopt the Work Programme and the Budget;
- (e) monitor the Strategic Plan, the Work Programme and the Budget and make adjustments as needed;
- (f) decide the timing and scope of independent reviews of GBIF, implementation, governance, impact or uptake;
- (g) adjust, by consensus, the scales of financial contributions suggested in Annex I, using appropriate economic indicators such as GDP;
- (h) adopt such rules, regulations and policies as may be required for the operations of GBIF;
- (i) monitor the performance of the GBIF Secretariat Host; if necessary, the Governing Board may replace the GBIF Secretariat Host;
- (j) select the Executive Secretary; the Governing Board may also remove the Executive Secretary;
- (k) approve the Staff Rules for the GBIF Secretariat based on recommendations from the Executive Secretary;
- (l) provide guidance and direction to the Executive Secretary on the duties of the position and monitor the Executive Secretary’s performance;
- (m) approve the annual financial statement and select the audit company;
- (n) carry out the other functions conferred upon it by this MOU, including by any Annexes or modifications hereto; and
- (o) consider any matters pertaining to GBIF or its operations submitted to it by the Executive Secretary, the GBIF Secretariat Host, or by any Participant;

6. Executive Committee

The Governing Board may establish an Executive Committee that is responsible for monitoring the performance of the Secretariat in implementing the decisions of the Governing Board, including the Strategic Plan, the Work Programmes, and budgets, and for making those intersessional decisions delegated to it by the Governing Board.

Paragraph 5
THE GBIF SECRETARIAT HOST

1. Role and Purpose

The GBIF Secretariat Host will provide the location, facilities and services agreed to in an arrangement between the Governing Board and the GBIF Secretariat Host. The services may cover staff management, financial management, accountancy, legal assistance, etc. The GBIF Secretariat Host may house the GBIF Secretariat and manage it in accordance with the laws in force in the country of the GBIF Secretariat Host. The GBIF Secretariat Host will also obtain or provide legal status for the GBIF Secretariat.

2. Reimbursement of Costs

- (a) Through appropriate financial arrangements with the GBIF Secretariat, expenses and costs reasonably and properly incurred by the GBIF Secretariat Host in supporting the GBIF Secretariat, above those costs that the GBIF Secretariat Host itself has agreed to provide, may - based upon a prior arrangement by the Executive Secretary - be paid from the funds collected pursuant to Paragraph 9.
- (b) Neither the GBIF Secretariat Host, nor its experts, employees, agents, representatives or contractors are entitled to commit the Participants to any expenditure beyond what is available in the Central Fund, as defined in Paragraph 9.1(b).

Paragraph 6
THE GBIF SECRETARIAT

1. Role and Responsibility

The GBIF Secretariat will execute the Work Programme in accordance with the Strategic Plan, and spend the Budget, under the direction of the Executive Secretary.

2. Designation

The GBIF Secretariat will consist of the Executive Secretary and such other staff as are judged necessary by the Governing Board.

3. Accountability

The GBIF Secretariat will be accountable through the Executive Secretary to the Governing Board for the execution of all scientific, financial and administrative activities undertaken to implement the GBIF Work Programme. The activities of the GBIF Secretariat will be subject to the laws and jurisdictions in force in the country of the GBIF Secretariat Host.

4. Tasks

The GBIF Secretariat will:

- (a) employ the Executive Secretary and other GBIF Secretariat staff;
- (b) be the holder of the Central Fund described in Paragraph 9.1;
- (c) be responsible for developing financial arrangements with Voting Participants specifying how those Participants will make their financial contributions to the Central Fund; and
- (d) hold in trust, and for the benefit of the Participants, all assets which may accrue to or be acquired for GBIF.

5. Transfer of Tasks to the Secretariat Host

Through appropriate financial arrangements between the Secretariat Host and the Secretariat, and with the approval of the Governing Board, some or all of the tasks listed in Paragraph 6.4 may be transferred to the GBIF Secretariat Host.

Paragraph 7

THE EXECUTIVE SECRETARY

1. Role and Authority

- (a) The Executive Secretary will act as the chief executive officer of GBIF and the Director of the GBIF Secretariat.
- (b) The Executive Secretary will have the authority, within limits and guidelines decided by the Governing Board, and, subject to the provisions of this Memorandum of Understanding, to enter into contracts and administer funds on behalf of GBIF.
- (c) The activities of the Executive Secretary will be subject to the laws and jurisdictions in force in the country of the GBIF Secretariat Host.

2. Accountability

The Executive Secretary will be accountable to the Governing Board for all scientific, financial and administrative activities of the GBIF Secretariat.

3. Responsibility

The responsibilities of the Executive Secretary include:

- (a) implementing the Work Programme and expenditure of the Budget;
- (b) hiring, such staff as may be required to carry out the Work Programme;
- (c) supervising the work of the GBIF Secretariat and its staff, including consultants and seconded personnel;
- (d) preparing and submitting to the Governing Board, not later than three months before the beginning of each financial year, a draft annual Work Programme and a Budget, together with an indicative Draft Work Programme and a Draft Budget for the following two years;
- (e) providing the Governing Board with a technically substantive Annual Report and an audited Financial Statement for approval; and
- (f) representing GBIF as appropriate.

Paragraph 8
INTELLECTUAL PROPERTY

1. Applicable Law

Nothing in this MOU should be read to alter the scope and application of Intellectual Property Rights and benefit sharing agreements as determined under relevant laws, regulations and international agreements of the Participants.

2. Access to Data

To the greatest extent possible, GBIF is an open-access facility. All users, whether GBIF Participants or others, should have equal access to data in databases affiliated with or developed by GBIF.

3. Intellectual Property Rights to Biodiversity Data

GBIF promotes the free dissemination of biodiversity data and, in particular:

- (a) should not assert any proprietary rights to the data in databases that are developed by other organisations and that subsequently become affiliated to GBIF;
- (b) should seek, to the greatest extent possible, to make freely and openly available, with the least possible restrictions on reuse, any data commissioned, created or developed directly by GBIF; and
- (c) should respect conditions set by data providers that affiliate their databases to GBIF.

When establishing affiliations or linkages with other databases, GBIF should seek to ensure that the data so made available will not be subject to limitations on the further non-commercial use and dissemination of those data, apart from due attribution of their source.

4. Attribution

GBIF should seek to ensure that the source of data is acknowledged and should request that such attribution be maintained in any subsequent use of the data.

5. Access to Specific Data

Nothing in this MOU should be read to restrict the right of owners of databases affiliated with GBIF to block access to any data.

6. Validity of Data

It should be a condition of access to and use of GBIF that users acknowledge that the validity of the data in any databases affiliated with GBIF cannot be assured. GBIF should disclaim responsibility for the accuracy and reliability of the data as well as for the suitability of its application for any particular purpose.

7. Legitimacy of Data Collection

Where the collection of new data has entailed access to biodiversity resources, GBIF should ask for reasonable assurances from the data provider that such access was consistent with applicable laws, regulations and any relevant requirements for prior informed consent.

8. Intellectual Property Rights to Biodiversity Tools

GBIF may claim appropriate Intellectual Property Rights available within applicable national jurisdictions over any tools, such as search engines or other software products that are developed by GBIF while carrying out the GBIF Work Programme.

9. Technology Transfer

The Participants acknowledge that, subject to any relevant Intellectual Property Rights, GBIF should seek to promote the non-exclusive transfer, on mutually agreed terms, to research institutions, particularly in developing countries, of such informatics technology as it has available, especially in conjunction with training and capacity development programs.

Paragraph 9

FINANCE

1. Basic Financial Contributions

- (a) Financial contributions made by Voting Participants in accordance with the scales in Annex I (and transferred to the GBIF Secretariat via the financial arrangements described in Paragraph 10.2) are considered to be Basic Financial Contributions.
- (b) These contributions are to be held by the GBIF Secretariat in a Central Fund and used as stipulated in the Budget approved by Governing Board.
- (c) The scales for Basic Financial Contributions are to be reviewed and approved by the Governing Board every five years, unless the scales are changed by consensus of the Governing Board

2. Supplementary Financial Contributions

- (a) In addition to Basic Financial Contributions, both Voting and Associate Participants may make Supplementary Financial Contributions to fund specific parts of the Work Programme, or for other specified purposes agreed to by the Governing Board. Those specified purposes may include facilitating attendance by Participants from developing countries at meetings of the Governing Board.
- (b) Supplementary Financial Contributions are to be held by the GBIF Secretariat, kept separate from other contributions, and used only for the purposes specified by the Participants making them.

3. Other Income

The Secretariat may accept other income from additional sources (e.g. foundations, agencies, research councils, and private companies) offered for the purposes set out in this MOU and the Strategic Plan.

4. Costs Borne by Participants

Participants bear the costs of their own participation in GBIF, including the costs of establishing and supporting their Participant Node(s), formulating or transmitting reports, travel costs for their delegates, and other expenses related to attendance by their representatives at meetings of the Governing Board and other GBIF functions, events, and activities.

5. Crediting of Income

Any income generated in the course of GBIF activities that accrues to the GBIF Secretariat is to be used for advancing the GBIF Work Programme.

Paragraph 10

ASSOCIATION AND DISASSOCIATION OF PARTICIPANTS

1. Association of Participants

Association with this MOU is open to any country, economy, inter-governmental or international organisation or to an entity designated by them. Such association becomes effective upon signature of this MOU.

2. Participant Status

- (a) A country that has signed the MOU-becomes eligible to be a Voting Participant on the Governing Board by making the financial arrangement negotiated with the GBIF Secretariat, as described in Paragraph 6.4(c). This arrangement sets out the Voting Participant's financial contribution as suggested in Annex I, and the period for which the arrangement is valid.
- (b) In order to retain its voting status, a Voting Participant must make its financial contribution each year within six months of the due date set out in the financial arrangement described in Paragraph 6.4(c).

3. Disassociation of Participants

- (a) Any Participant may disassociate itself from this MOU by advising the Governing Board in writing of its intention to do so and of the effective date.
- (b) In the event of disassociation of a Voting Participant, the Governing Board may decide by consensus to adjust the Work Programme and the Budget to take account of such disassociation or, again by consensus, may decide to adjust the scale of contributions of Participants to the Budget.

Paragraph 11
OTHER MATTERS

1. Duration

(a) GBIF is intended to be a long-term cooperative endeavour, in order to sustain the benefits of access to biodiversity data.

(b) This MOU covers the period 1 January 2007 to 31 December 2011.

2. Termination

The Voting Participants, acting by consensus, may terminate this MOU at any time. In a situation where termination or expiration of this MOU occurs without a new MOU or other document being in place, the GBIF Secretariat, acting in accordance with the laws of the jurisdiction in which it is located, will arrange for the liquidation of the assets of GBIF; property held by the GBIF Secretariat for the benefit of the Participants is to be regarded, for this purpose, as assets of GBIF. In the event of such liquidation, the GBIF Secretariat, so far as practicable, will distribute any assets of GBIF, or the proceeds therefrom, in proportion to the basic financial contributions which the Participants have made from the beginning of the operation of GBIF.

3. Annexes

Annexes to this MOU are an integral part of the document.

4. Modifications

Excepting paragraph 2 (5), this MOU (including its Annexes) may be modified at any time by the Governing Board through a consensus vote of all the representatives of all the Voting Participants present and voting at the yearly business meeting of the Governing Board.

Signed at _____ this ____ day of _____, 20__.

Annex I

FINANCIAL CONTRIBUTIONS FOR VOTING PARTICIPANTS

1. Intent of this Annex

This Annex describes the suggested financial contributions for voting participation in GBIF for the five financial years: 1 January 2007 to 31 December 2011.

2. Classes of Voting Participants

For purposes of the financial contributions, there are two classes of Voting Participants. "Existing Voting Participants" are defined as the 26 Voting Participants which were in place in the last year of the previous MOU (2006). "New Voting Participants" are those which did not have Voting Participant status in the first MOU.

3. Suggested Basic Financial Contributions

The criteria for the calculation of the basic financial contribution for Voting Participants and the actual amount to be paid in each financial year are listed in Table 1 below.

3.1 For Existing Voting Participants, the GDP figures for 2004, as listed on the World Bank web-site, determine the basic contribution, with the proviso that during any year of this MOU, no Existing Voting Participant shall contribute a sum lower than its contribution under the first GBIF MOU (2001-2006).

3.2 For New Voting Participants, the contribution is based on the country's most recent GDP as listed on the World Bank web-site, at the time they become a Voting Participant.

4. Minimum Contribution

The minimum contributions to GBIF will be 500 Euros.

5. Cap

The basic contribution for any country is capped at 23.5% of the total core budget.

6. Reduction in contributions for countries with a per capita GDP less than 10.000 USD.

Countries whose per capita GDP according to the World Bank country statistics is less than 10.000 USD may contribute with 50% of the required amount, but never below the minimum contribution.

7. New Voting Participant

The basic financial contribution for a New Voting Participant is determined by its GDP and the year in which it becomes a Voting Participant.

However, for the first year of its participation in GBIF, a New Voting Participant may acquire voting rights by making a contribution of at least one half of the suggested amount, as long as the contribution does not go below the minimum contribution of 500 Euros.

Contributions for subsequent years shall be at the full level in order to retain voting rights.

8. Fixed contribution for 2007-2011

Once a country becomes a Voting Participant or continues under this MOU to be a Voting Participant, its basic financial contribution will be fixed for the duration of this MOU at the amounts shown in, or derived from, Table 1 below, even if its GDP subsequently changes.

9. Addition or departure of Voting Participants

Neither the addition of a New Voting Participant, nor departure of a Voting Participant, will affect the contributions of the remaining members, unless the Governing Board decides to change the payment structure as allowed under Paragraph 4.5(g) of this MOU.

10. Negotiation of alternative payment schedules

When negotiating the financial arrangement with a Voting Participant, the Secretariat, with approval of the Executive Committee, may accept a payment schedule of the basic financial contribution that deviates from Table 1.

11. Payment of contributions

The suggested basic financial contribution may be paid either in Euros or in the equivalent amount of Danish Kroner (DKK) unless another currency is accepted by the Executive Secretary in a financial arrangement as described in Paragraph 10.2 of this MOU.

Details of the financial contributions will be specified in the GBIF Financial Regulations.

Suggested Annual Basic Financial Contribution for 2007-2011

The basic financial contribution from each Voting Participant is proportional to its GDP, applying the modifications listed in points 3-6 above.

Table 1 shows the suggested basic financial contribution for Existing Voting Participants, taking into account provisions 3-6 above.

To determine the basic financial contribution for a New Voting Participant, look up the country's latest GDP as listed at the World Bank web site, and use that amount (in billions) to determine the suggested basic financial contribution for the year of becoming a Voting Participant, following the instructions in Table 2.

Table 1:**Suggested basic financial contributions for Existing Voting Participants (in Euros)**

Current Voting Participants	2007 contribution	2008 contribution	2009 contribution	2010 contribution	2011 contribution	Average contribution rounded
United States of America	646.300	743.200	819.700	887.800	947.300	808.900
Japan	556.600	640.100	706.000	764.600	815.800	696.600
Germany	326.700	375.700	414.400	448.800	478.900	408.900
United Kingdom	254.100	292.200	322.300	349.100	372.500	318.000
France	242.300	278.600	307.300	332.800	355.100	303.200
Spain	119.900	137.900	152.100	164.700	175.700	150.100
Canada	118.500	136.300	150.300	162.800	173.700	148.300
Republic of Korea	82.200	94.600	104.300	113.000	120.500	102.900
Australia	79.500	87.900	97.000	105.000	112.100	96.300
Netherlands	79.500	80.300	88.600	96.000	102.400	89.400
Belgium	79.500	79.500	79.500	79.500	79.500	79.500
Sweden	79.500	79.500	79.500	79.500	79.500	79.500
Denmark	79.500	79.500	79.500	79.500	79.500	79.500
Norway	79.500	79.500	79.500	79.500	79.500	79.500
Finland	79.500	79.500	79.500	79.500	79.500	79.500
Portugal	79.500	79.500	79.500	79.500	79.500	79.500
Mexico	39.900	45.900	50.600	54.800	58.400	49.900
South Africa	40.000	40.000	40.000	40.000	40.000	40.000
New Zealand	40.000	40.000	40.000	40.000	40.000	40.000
Peru	4.100	4.700	5.200	5.700	6.000	5.100
Slovenia	3.900	4.400	4.900	5.300	5.600	4.800
Costa Rica	1.100	1.300	1.400	1.500	1.600	1.400
Iceland	1.400	1.600	1.700	1.900	2.000	1.700
Estonia	550	630	700	760	810	690
Nicaragua	500	500	500	500	500	500
Equatorial Guinea	500	500	500	500	500	500

Table 2:**Calculating the suggested basic financial contributions for New Voting Participants (in Euros)**

To calculate the suggested financial contribution, go to the World Bank web site (www.worldbank.org) to get the latest GDP information for your country.

Use the following table to calculate the basic financial contribution for the year in which you are to become a Voting Participant, and all subsequent years.

To calculate the basic financial contribution for each year, multiply the GDP (in billions of USD) by the multiplier

Year	2007	2008	2009	2010	2011
Multiplier	121,00	139,15	153,48	166,22	177,36

This gives the *unadjusted* requested financial contribution, in Euros, for the years indicated.

Then apply modifications 4-7 from the text above if any of them apply to your country.

INTELLECTUAL PROPERTY

1. **Applicable Law**

Nothing in this MOU should be read to alter the scope and application of Intellectual Property Rights and benefit sharing agreements as determined under relevant laws, regulations and international agreements of the Participants.

2. **Access to Data**

To the greatest extent possible, GBIF is an open-access facility. All users, whether GBIF Participants or others, should have equal access to data in databases affiliated with or developed by GBIF.

3. **Intellectual Property Rights to Biodiversity Data**

GBIF promotes the free dissemination of biodiversity data and, in particular:

- a. should not assert any proprietary rights to the data in databases that are developed by other organisations and that subsequently become affiliated to GBIF;
- b. should seek, to the greatest extent possible, to make freely and openly available, with the least possible restrictions on reuse, any data commissioned, created or developed directly by GBIF; and
- c. should respect conditions set by data providers that affiliate their databases to GBIF.

When establishing affiliations or linkages with other databases, GBIF should seek to ensure that the data so made available will not be subject to limitations on the further non-commercial use and dissemination of those data, apart from due attribution of their source.

4. **Attribution**

GBIF should seek to ensure that the source of data is acknowledged and should request that such attribution be maintained in any subsequent use of the data.

5. **Access to Specific Data**

Nothing in this MOU should be read to restrict the right of owners of databases affiliated with GBIF to block access to any data.

6. **Validity of Data**

It should be a condition of access to and use of GBIF that users acknowledge that the validity of the data in any databases affiliated with GBIF cannot be assured. GBIF should disclaim responsibility for the accuracy and reliability of the data as well as for the suitability of its application for any particular purpose.

7. **Legitimacy of Data Collection**

Where the collection of new data has entailed access to biodiversity resources, GBIF should ask for reasonable assurances from the data provider that such access was consistent with applicable laws, regulations and any relevant requirements for prior informed consent.

8. **Intellectual Property Rights to Biodiversity Tools**

GBIF may claim appropriate Intellectual Property Rights available within applicable national jurisdictions over any tools, such as search engines or other software products that are developed by GBIF while carrying out the GBIF Work Programme.

9. **Technology Transfer**

The Participants acknowledge that, subject to any relevant Intellectual Property Rights, GBIF should seek to promote the non-exclusive transfer, on mutually agreed terms, to research institutions, particularly in developing countries, of such informatics technology as it has available, especially in conjunction with training and capacity development programs.

Promoting Access to Public Research Data for Scientific, Economic, and Social Development

P Arzberger,^{1*} P Schroeder,² A Beaulieu,³ G Bowker,⁴ K Casey,¹ L Laaksonen,⁵ D Moorman,⁶ P Uhler,⁷ P Wouters³

¹University of California San Diego, 9500 Gilman Dr., San Diego, California, USA, Email: parzberg@ucsd.edu, kcasey@ucsd.edu

²Ministry of Education, Culture and Science, The Netherlands, Email: p.schroeder@minocw.nl

³Networked Research and Digital Information (Nerdi), The Netherlands, Email: Anne.Beaulieu@niwi.knaw.nl, Paul.Wouters@niwi.knaw.nl

⁴Santa Clara University, Santa Clara, California, USA, Email: gbowker@scu.edu

⁵CSC - Scientific Computing Ltd, Finland, Email: Leif.Laaksonen@csc.fi

⁶Social Sciences and Humanities Research Council, Canada, Email: DAVID.MOORMAN@SSHRC.CA

⁷U.S. National Academy of Sciences/National Research Council, USA, Email: puhler@nas.edu

ABSTRACT

Access to and sharing of data are essential for the conduct and advancement of science. This article argues that publicly funded research data should be openly available to the maximum extent possible. To seize upon advancements of cyberinfrastructure and the explosion of data in a range of scientific disciplines, this access to and sharing of publicly funded data must be advanced within an international framework, beyond technological solutions. The authors, members of an OECD Follow-up Group, present their research findings, based closely on their report to OECD, on key issues in data access, as well as operating principles and management aspects necessary to successful data access regimes.

Keywords: Data access, Science policy, Data sharing, Data management, Database, Archives, Scientific infrastructure, Global e-science, OECD, Public domain

1 INTRODUCTION AND SUMMARY

It is now commonplace to say that information and communications technologies are rapidly transforming the world of research. We are only beginning to recognize, however, that management of the scientific enterprise must adapt if we, as a society, are to take full advantage of the knowledge and understanding generated by researchers. One of the most important areas of information and communication technology (ICT)-driven change is the emergence of e-science, briefly defined as increased access, via desktop or other interface via the Internet, to distributed resources, global collaboration, and the intellectual, historical, analytical, and investigative output of a range of scientific communities (Atkins, Droegemeier, Feldman, Garcia-Molina, Klein, Messerschmitt, et al., 2003; Research Councils UK, n.d.).

In recent years, the debate on e-science has tended to focus on the “open access” to the digital *output* of scientific research, namely, the results of research published by researchers as the articles in the scientific journals (Access all Areas, 2004; for recent discussions on open access see Cook (2004), Suber (2004) and House of Commons Science and Technology Committee (2004)). This focus on publications often overshadows the issues of access to the *input* of research - the research data, the raw material at the heart of the scientific process and the object of significant annual public investments. In terms of access, availability of research data generally poses more serious problems than access to publications.

Ensuring research data are easily accessible, so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources. Moreover, as research becomes increasingly global, there is a growing need to systematically address data access and sharing issues beyond national jurisdictions. The goals of this article and its recommendations are to ensure that both researchers and the public receive optimum returns on the public investments in research, and to build on the value chain of investments in research and its data resource.

To some extent, research data are shared today, often quite extensively within established networks, using both the latest technology and innovative management techniques. The Follow Up Group, which is identified in the Acknowledgments section of this paper, drew on the experiences of several of these networks to examine the roles and responsibilities of governments as they relate to data produced from publicly funded research. The objective was to seek good practices that can be used by national governments, international bodies, and scientists in other areas of research. In doing so, the Group developed an analytical framework for determining where further improvements can be made in the national and international organization, management, and regulation of research data (Arzberger, Schroeder, Beaulieu, Bowker, Casey, Laaksonen, et al., 2004).

The findings and recommendations presented here are based on the central principle that ***publicly funded research data should be openly available to the maximum extent possible***. Availability should be restricted only by legitimate considerations of national security restrictions; protection of confidentiality and privacy; intellectual property rights; and time-limited exclusive use by principal investigators. Publicly funded research data are a public good, produced in the public interest. As such they should remain in the public realm. This does not preclude the subsequent commercialization of research results in patents and copyrights, or of the data themselves in databases, but it does mean that a copy of the data must be maintained and made openly accessible. Implicitly or explicitly, this principle is recognized by many of the world's leading scientific institutions, organizations, and agencies. Expanding the adoption of this principle to national and international stages will enable researchers, empower citizens and convey tremendous scientific, economic, and social benefits.

Evidence from the case studies and from other investigations undertaken for this report suggest that successful research data access and sharing arrangements, or regimes, share a number of key attributes and operating principles. These bring effective organization and management to the distribution and exchange of data. The key attributes include: openness; transparency of access and active dissemination; the assignment and assumption of formal responsibilities; interoperability; quality control; operational efficiency and flexibility; respect for private intellectual property and other ethical and legal matters; accountability; and professionalism. Whether they are discipline-specific or issue oriented, national or international, the regimes that adhere to these operating principles reap the greatest returns from the use of research data.

There are five broad groups of issues that stand out in any examination of research data access and sharing regimes. The Follow Up Group used these as an analytical framework for examining the case studies that informed this report, and in doing so, came to several broad conclusions:

- Technological issues: Broad access to research data, and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls;
- Institutional and managerial issues: While the core open access principle applies to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers;
- Financial and budgetary issues: Scientific data infrastructure requires continued, and dedicated, budgetary planning and appropriate financial support. The use of research data cannot be maximized if access, management, and preservation costs are an add-on or after-thought in research projects;
- Legal and policy issues: National laws and international agreements directly affect data access and sharing practices, despite the fact that they are often adopted without due consideration of the impact on the sharing of publicly funded research data;
- Cultural and behavioural issues: Appropriate reward structures are a necessary component for promoting data access and sharing practices. These apply to those who produce and those who manage research data.

The case studies and other research conducted for this report suggest that concrete, beneficial actions can be taken by the different actors involved in making possible access to, and sharing of, publicly funded research data. This includes the Organization for Economic Cooperation and Development (OECD) as an international organization with credibility and stature in the science policy area. At the March 2003 meeting of the OECD Committee of Science and Technology Policy, the Follow Up Group recommended that the OECD consider the following:

- Put the issues of data access and sharing on the agenda of the next Ministerial meeting (see Section 7, and Declaration on access to public research data from public funding (2004));
- In conjunction with relevant member country research organizations,

- Conduct or coordinate a study to survey national laws and policies that affect data access and sharing practices;
- Conduct or coordinate a study to compile model licensing agreements and templates for access to and sharing of publicly funded data;

With the rapid advances in scientific communications made possible by recent developments in ICTs, there are many aspects of research data access and sharing that have not been addressed sufficiently by this report, would benefit from further study, and will need further clarification. Accordingly, further possible actions that could be considered include:

- Governments from OECD expand their policy frameworks of research data access and sharing to include data produced from a mixture of public and private funds;
- OECD consider examinations of research data access and sharing to include issues of interacting with developing countries; and
- OECD promote further research, including a comprehensive economic analysis of existing data access regimes, at both the national and research project or program levels.

National governments have a crucial role to play in promoting and supporting data accessibility since they provide the necessary resources, establish overall policies for data management, regulate matters such as the protection of confidentiality and privacy, and determine restrictions based on national security. Most importantly, national governments are responsible for major research support and funding organizations, and it is here that many of the managerial aspects of data sharing need to be addressed. Drawing on good practices worldwide, the Follow Up Group suggests that national governments should consider the following:

- Adopt and effectively implement the principle that data produced from publicly funded research should be openly available to the maximum extent possible;
- Encourage their research funding agencies and major data producing departments to work together to find ways to enhance access to statistical data, such as census materials and surveys;
- Adopt free access or marginal cost pricing policies for the dissemination of research-useful data produced by government departments and agencies;
- Analyze, assess, and monitor policies, programs, and management practices related to data access and sharing policies within their national research and research funding organizations.

The widespread national, international, and cross-disciplinary sharing of research data is no longer a technological impossibility. Technology itself, however, will not fulfill the promise of e-science. Information and communication technologies provide the physical infrastructure. It is up to national governments, international agencies, research institutions, and scientists themselves to ensure that the institutional, financial and economic, legal, and cultural and behavioural aspects of data sharing are taken into account.

2 BACKGROUND TO THE STUDY

At its March 2001 meeting, the OECD Committee on Scientific and Technology Policy (CSTP) accepted a proposal from The Netherlands to establish a working group on issues of access to research information. The plans of the working group were presented at the October 2001 CSTP meeting. Subsequently, the Committee narrowed the scope of activities to access to and sharing of research data produced from public funding. Participation in the group was broadened to include Australia, Canada, Denmark, Finland, Germany, Japan, Poland, the Netherlands, the United Kingdom, and the United States. The CSTP asked the working group to:

- Report on current practices concerning access to and sharing of research data and their underlying principles on the basis of case studies;
- Report on the effects of selected current data sharing practices on the quality of research and the progress of science;
- Suggest principles for making policy on data sharing within the relevant national and international policies and regulatory frameworks.

The report's core principle is that *publicly funded research data should be openly available to the maximum extent possible*. Adoption of this principle will promote good stewardship of public knowledge, develop strong value chains of innovation, and maximize benefits from international cooperation (see Box 1). The report's findings and recommendations are addressed to: CSTP members as representatives from the governments of OECD member countries; and professional and scholarly associations. The objective is to contribute to a better understanding of the importance of research data access and sharing, and to offer suggestions on how the new digital challenges should be met.

Building on a number of case studies and a great deal of other research, the report focuses on issues related to the access and sharing of publicly funded research data, in digital form, across all disciplines in the natural, health, and social sciences (Wouters & Schröder, 2003). Attention is paid to the international aspects of access and sharing relevant to scientific cooperation among OECD member states. Three significant topical areas fell outside the charge of this working group, however, and will require separate follow-up: issues particular to developing countries; issues related to data produced by a mixture of public and private funding; and the issue of national security restrictions in light of recent global events since 11 September 2001 (on issues of national security and data access, see Mervis & Stokstad (2002)).

Box 1: This core principle guides many public scientific institutions and scientists. However, it remains unevenly implemented. Most recently, it was adopted by the United Kingdom's Medical Research Council. After a workshop hosted by the European Science Foundation, the MRC drafted the following statement: MRC promotes the creation of a diverse range of datasets, many of which are rich in informational content, unique and cannot be readily replicated. Sharing allows scientists to extend the value of these datasets through new, high quality, ethical research and exploitation. It also reduces unnecessary duplication of data collection. Building preservation systematically into routine data management is part of good research practice: it strengthens quality, enables replication and audit, and provides a sound basis for data sharing Medical Research Council (n.d.).

In this report, we define "access to data" as the act of making the data available for use by others; by "sharing" we mean a researcher allowing one or more other individuals to use data, typically with the implicit, if not explicit assumption that it is on a reciprocal basis. The sharing of data involves providing specific access, whereas the act of providing access by itself does not necessarily involve any sharing arrangement. In this article, data sharing focuses on data exchanges between individuals or groups of researchers rather than institutions, while access may be provided at any level. Sharing in our view also reflects the cooperative norms of public science as practiced within many disciplines by many researchers in OECD countries. The U.S. National Institutes of Health on the other hand, use the term "data sharing" throughout all of their formal, institutional regulations on the use of research data generated by NIH funding (NIH Office of Extra Mural Research, n.d.). We define data as in the U.S. National Institutes of Health definition of final research data: "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings". In the OECD documents on the subject, "data are defined as the factual records (numerical scores, textual records, images and sounds) used as sources and base material for scientific research." For other definitions of terms involved with data, see Westbrook & Grattidge (1991), National Research Council (1997) and Esanu & Uhler (2004).

3 KEY ISSUES IN DATA ACCESS AND DATA SHARING

3.1 The changing information technology context for scientific research and innovation

Information and communication technologies (ICTs) are enabling the rapid transformation of an increasing number of research areas as well as the broader society: witness the growth in the number of Internet hosts per person, in the percentage of computers per household, and in the continued rate of growth of chip, storage, and network technology capacity (Stix, 2001). Concurrently, there has been an explosion in the amount of data produced across all types of scientific endeavour. Examples of an this explosive increase in data production range from genetic sequence and protein structure data in bioinformatics, to various types of brain imagery in neuroscience, to sky surveys and virtual observatories in astronomy, and geospatial data such as Global Spatial Data Infrastructure.

Continuing ICT advances, such as the development of grid computing, large-capacity optical transmission networks, wireless networks of sensors and devices, and complex imaging systems, promise to push these transformations farther and faster. ICT-dependent research, such as geographic information systems, data visualisation systems, and realistic modelling, are adding tremendously to our ability to study and understand the world in which we live. These developments provide researchers in OECD countries, and increasingly in developing countries, with the opportunity not only to be more efficient, more effective and better connected, but also to dramatically expand the scope and nature of their investigations. This expansion of scope of scientific investigation results from activities such as combining data from multiple data sources to gain a greater statistical power to resolve hypotheses (for example, see Biomedical Informatics Research Network (n.d.)) and obtaining real-time global measurement on environmental observations. Together they create the possibility of an “e-science infrastructure.” The growing activities in data collection, storage, processing, distribution, and preservation are, however, only loosely connected. They require systematic planning to realize the full potential of the emerging e-science infrastructure.

3.2 The benefits of data access and sharing in public research

Within this new technological context, more widespread and efficient access to and sharing of research data can be expected to have substantial benefits for public scientific research (see Box 2). Open access to, and sharing of, data reinforces open scientific inquiry, encourages diversity of analysis and opinion, promotes new research, makes possible the testing of new or alternative hypotheses and methods of analysis, supports studies on data collection

BOX 2: Access to international data has helped produce a better understanding of public health issues and worldwide disease prevention and control. For instance, research on cholera outbreaks and their relationship to numerous environmental factors relied upon data drawn from epidemiology, NASA remote sensing, marine biology, microbiology, genomic data, and social science data. This research—an example of ‘biocomplexity’ studies supported by the U.S. National Science Foundation—would have been impossible without access to numerous databases. The effect of this interdisciplinary and international research project is an increased scientific and sociological understanding of cholera outbreaks and their prevention (Colwell, 2002).

methods and measurement, facilitates the education of new researchers, enables the exploration of topics not envisioned by the initial investigators, and permits the creation of new data sets when data from multiple sources are combined. Sharing and open access to publicly funded research data not only helps to maximize the research potential of new digital technologies and networks, but provides greater returns from the public investment in research (Fienberg, Martin, & Staf, 1985; National Research Council, 1999).

Moreover, improving and expanding the open availability of public research data will generate wealth through the downstream commercialisation of outputs, provide decision-makers with the necessary facts to address complex, often trans-national problems, and offer individuals the opportunity to better understand the social and physical world in which we all live. For example, a recent analysis demonstrated the economic benefits of providing open access to government meteorological data without any restrictions on re-use (Weiss, 2003; Weiss, 2002; European Union Green Paper, 1998; PIRA International, 2000). The “value adding” meteorological information industry in the United States has revenues in excess of \$500M annually. The public meteorological data also support a rapidly growing weather risk management industry that underwrites financial risk management instruments valued at approximately \$8B. In contrast, the private-sector value adding industry for meteorological information in the European Union is very small, largely attributable to the highly restrictive data policies of most national governmental meteorological services. What are harder to measure, but certainly occur, are the lost opportunities for researchers, students, and various other potential public users who find the high costs of the E.U.’s public data to be too great to access and use.

As a key link in the value chain of investments in research, open access to factual data plays an increasingly important role in all these areas.

3.3 Roles and responsibilities of governments

If researchers throughout the world are to take full advantage of ICTs to improve and expand access to, and sharing of, research data, existing technological, institutional and managerial, financial and budgetary, legal and policy, and

BOX 3: Poor stewardship and lost opportunity costs for data access is exemplified by the case of Statistics Canada, which attempted to recover costs for its data management by charging data users. The effect of this form of management of these public data was a dramatic decrease in their use. In a study of the case, it was found that “Cost recovery was supposed to introduce a market type discipline on the demand for and supply of goods and services provided by the government. Since in economic terms Statistics Canada’s outputs are public goods, the type of discipline envisioned by this policy is impossible to attain. Instead we have users who complain, refuse to pay and generally attempt to find alternative sources for their information needs. This policy fails the improved management of resources test (McMahon, 1996)”

cultural and behavioural aspects must be addressed comprehensively and in an integrated way. To date, these aspects have often been treated on an ad hoc, project-specific basis. Given that OECD countries spend tens of billions of dollars each year collecting data that can be used for research and for other social and economic benefits, ensuring that these data are easily accessible so that they can be used as often and as widely as possible, is a matter of sound stewardship of public resources (see Box 3).

Scientists, research institutions, and research funding agencies around the world are increasingly engaging in large-scale, data-intensive projects. Such projects require data-management infrastructure, data-exchange protocols and policy frameworks, and a broad professional understanding that more extensive availability and use of the data is both necessary and desirable. Over

the past decade, numerous studies, disciplines, research programs, and agencies have begun to address the complexities and benefits of open data access and sharing arrangements (National Research Council, 1997; Medical Research Council, n.d.). As scientists become better connected with each other, particularly through the Internet, and as research focuses on issues of global importance, such as climate change, human health and biodiversity, there is growing need to systematically address data access and sharing issues beyond national jurisdictions and thereby create greater value from international co-operation. The goal should be to ensure that both researchers and the broader public receive the optimum return on public investments, and to build on the value chain of investments in research and research data (Stiglitz, Orszag & Orszag, 2000).

4 CORE PRINCIPLE AND PREMISES

The findings and recommendations that follow are based on the central principle that *publicly funded research data should be openly available to the maximum extent possible*.

As a general principle, publicly funded research data should be as open as possible and available at the lowest possible access cost, subject only to legitimate restriction and considerations. Restrictions may be necessary for reasons of national security, for the protection of privacy of citizens, or the confidentiality of trade secrets. Access to and use of research data may be limited by the respect for private intellectual property rights. Finally, there may be reasons for granting periods of temporary exclusive use to those who collected the data. But the guiding principle should be openness.

In order to derive the maximum benefit from public investments in research data, access, use, management and preservation must be an integral part of the research process. Conversely, data should not be considered an expendable by-product of research. In many cases, data have value beyond the project and anticipated use for which they were originally collected. The re-use of publicly funded data for research and other types of applications should be promoted and not restricted.

The accessing and sharing of data is not merely a technical matter, but also a complex social process in which researchers have to balance different pressures and interests. Purely regulatory approaches to data sharing are not likely to be successful without consideration of these factors. Various approaches to data access and sharing are therefore necessary, including the establishment of regulations and incentives, and the dissemination of best practices.

The following three premises complement and support the core principle of this report:

4.1 Data from publicly funded research are a public good produced in the public interest

Both the data from publicly funded research and research itself have strong public good characteristics, as elaborated by Kaul, Grunberg & Stern (1999), that support their open availability to the public, and especially to other researchers.

4.2 Factual data are central to the scientific research process

The production, open dissemination, and unfettered use of factual data are essential attributes of, and inputs to, modern systems of scientific research and technological innovation. Recognizing the role of digital data as fundamental to the value chain of science, technology, and innovation will enable an optimum return on public investments.

4.3 Data access and sharing issues are international in scope

To more fully exploit the possibilities of global digital networks, and to capture their benefits for the global community, policy issues concerning access to and sharing of publicly funded scientific research data must be addressed, not only at the institutional and national levels, but also at the international level.

5 DATA ACCESS OPERATING PRINCIPLES AND ATTRIBUTES

Data access and sharing requires effective organization and management. The necessary components that make up this organization and management may be characterized as “data access regimes.” In their ideal form, these regimes enable all participants in the scientific research process to freely and efficiently access and share data. Adequate data access regimes may require both distributed and centralised responsibilities across different management domains that include the technological, institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural.

Although no single approach to developing an effective data access regime is possible, a list of operating principles for and attributes of effective data access regimes and resources can be offered. This list of attributes and operating principles is based on a broad set of experiences, and supported by the case studies conducted for this article. The operating principles evolved out of recommendations developed by Franken (2001). Key attributes are listed below, and illustrated with an example from the case studies.

5.1 More explicit access regimes

There is a universal requirement for the formalisation of institutional rules and data management policies. The need for this formalisation follows from the growing complexity and scale of scientific research and the increasing expenditure on research data. At the moment, it is often not clear who is authorised to distribute data across the globe. To reach the necessary transparency in the tasks and responsibilities of those involved, terms of access to and use of data that rest on tacit agreements should be made explicit and formalised. A systematic and institutionalized approach is needed to help address operating characteristics of data access and to take advantage of the opportunities arising from publicly funded research.

5.2 Operating Principles

5.2.1 Openness

Open availability of publicly funded research data to the maximum extent possible is the core principle..

5.2.2 Transparency of access and active dissemination

Open data access requires actively disseminating where the data can be found, what the context and structure of the data collection is (metadata), how long the resource will be accessible, and what protocols and standards are employed. In short, this principle refers to the systematic visibility and traceability of data resources.

5.2.3 Assignment and assumption of formal responsibility

Formal responsibility for tasks associated with data access must be assumed by the appropriate participants in the global science system. The various individuals and institutions involved in the chain of data-related activities all have specific manifest and latent duties and obligations. These are founded in formal legal and professional normative standards and in the regulations of various agencies. Responsibility must also be assumed for various rights in the data supply, such as authorship, producer credits, ownership, financial arrangements, licensing terms, and, where appropriate, restrictions on use.

5.2.4 Professionalism

Codes of conduct, and related normative standards, of professional scientists and their communities can help to promote good practice and simplify the regulatory aspect of access regimes.

5.2.5 Interoperability

Technical and software standards and protocols are required to ensure the access and usability of data. These should be clear to the user and adopted by as many data management organizations as possible.

5.2.6 Quality

Quality refers to the proper description of uncertainties surrounding the production of the data (e.g., the techniques employed in their collection and archiving, and the measuring instruments and their calibration), the ability to ensure that the cited source and value are *authentic*, that the data retain *integrity* (complete and absent from introduced errors), and that they are *secure* against loss, destruction, modification, and unauthorized access.

5.2.7 Operational Efficiency

Open access to data increases the efficiency of research by avoiding unnecessary duplication of data collection and permitting the creation of new data sets by combining data from multiple sources. Coupled with open access, comprehensive documentation of data sets and how to access them provides a more efficient use of resources.

5.2.8 Flexibility

In general, scientific communities will approach data management requirements more consistently within their discipline internationally, than they will across other disciplines on a national level. Data access regimes need to be sufficiently flexible to take account of this variation.

5.2.9 Property

Institutional intellectual property rights as well as the individual rights of researchers are considerations of property interests. Unlike the private sector, public research operates on a principle of collective property interests, which are promoted by the open access and sharing of data resources.

5.2.10 Legality

Legal restrictions may limit access to and use of data. Examples of legal restrictions involve national security, privacy, and trade secrets. Restrictions will apply primarily to 'secondary' data sets compiled for purposes other than scientific research. In some cases, the sensitive parts of data sets can be left out without rendering them useless. Specific types of legal restrictions include: national security, privacy and the protection of trade secrets.

5.2.11 Accountability

Accountability involves measuring the cost, benefit, and performance of data access and sharing regimes and taking appropriate actions in response to the results.

5.3 Building a Data Access Regime: the Global Biodiversity Information Facility (GBIF)

The Global Biodiversity Information Facility (GBIF), which began under the auspices of the OECD Megascience Forum, has sought to implement these principles as a means to achieve the larger goal of providing worldwide access to biodiversity data. GBIF's goal is to make "the world's scientific biodiversity data freely available to all [openness]"(Global Biodiversity Information Facility, n.d.). The fundamental motivation for GBIF is to enable access to a vast amount of biodiversity data housed in databases distributed in numerous countries and institutions. By bringing all these data into one interoperable network, and producing a registry of biodiversity information resources, GBIF will produce systematic visibility and traceability of data resources [transparency].

Formal responsibilities of different participants involved in the task of building GBIF's organisation and legal relationships have been established in GBIF's Memorandum of Understanding. GBIF's Secretariat is responsible for carrying out work programmes that are approved by the Governing Board, which consists of representatives of GBIF's Participants. This structure enables GBIF to have a legal identity as an international body, and to manage financial contributions and work programmes, while drawing upon the additional separate efforts and resources of Participants. The establishment of GBIF's activities occurred through contact with existing scientific and political bodies to maintain and establish professional codes of conduct, gain consensus about scientific outcomes, and negotiate with government representatives about GBIF's larger social and economic roles [professionalism].

Participants will provide stable gateways, or "nodes," to databases that contain primary or meta-level biodiversity data. These nodes must provide documentation and metadata about the data in the databases, vouch for data **quality**, ensure data authenticity and security. GBIF will help develop standards for database **interoperability** through one of its 4 work programmes, Data Access and Database Interoperability (DADI). GBIF aims to develop an interoperable network of distributed databases by coordinating and leveraging existing national and international programs and projects, which allows for **operational efficiency** and more cost-effective basis for making biodiversity data freely and easily available to a heterogeneous user community.

The databases and the data accessed through GBIF are in most cases owned and developed by other organisations and thus will not entail any assertion of IPRs by GBIF itself [property]. GBIF intends to provide best practices on how to deal with IPRs, particularly since it will be drawing from databases hosted by different institutions and countries with different legal frameworks, with a view to promoting open access and sharing to the maximum extent possible. GBIF also asserts in its MOU that biodiversity data will be properly used and acknowledged by its participants [legality]. Further, its efforts are consistent with the Global Taxonomic Initiative of the Convention on Biological Diversity concerning the proper and equitable use of biodiversity data and the resources to which they refer.

During the establishment of GBIF, the OECD provided the forum to assess the level of support for this new scientific collaboration, to bring together related proposals and to develop detailed plans that could then be taken up by interested countries. According to paragraph 11.2 in GBIF's MOU, in the third year of its initial five-year period of existence, "an independent review of its operations, financial mechanisms, legal basis, governance structure, and links to other organizations will be conducted to determine if any changes are needed. The lessons learned will be used to evaluate the effectiveness of the governance structure and to recommend any necessary changes" [accountability]. That review is currently being conducted.

6 DATA ACCESS MANAGEMENT: FIVE DOMAINS

Efficient data access can only take place with the proper administration and organization of different management domains within data access regimes. These domains include technological, institutional and managerial, financial

and budgetary, legal and policy, and cultural and behavioural considerations (see Figure 1). The domains provide a framework for locating and analyzing where improvements to data access and sharing can be made.

The five domains differ in character across the traditions and practices of specific scientific disciplines, e.g., astrophysics, biology. Thus, data access regimes may vary in significant ways. There is no single model for how data access should take place. The implementation of the core principle of open availability, however, requires a systematic approach that recognizes the necessity of implementing improvements across the interdependent management domains. This approach also requires the involvement of actors from various levels: governments, funding agencies, research institutions and professional societies, as well as individual scientists themselves.

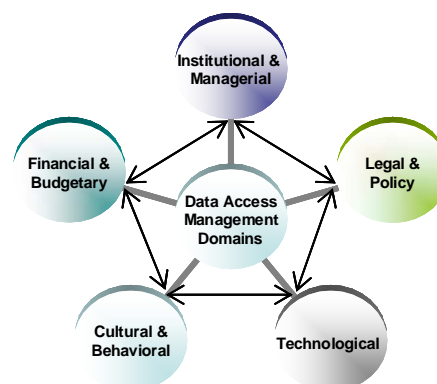


Figure 1. Components of a Data Access Regime

6.1 Technological domain

Broad access to research data, and their optimum exploitation, requires appropriately designed technological infrastructure, broad international agreement on interoperability, and effective data quality controls.

A technical infrastructure that supports user needs is necessary to derive maximum benefits from data access and sharing. This infrastructure must be robust enough for long-term use and, when appropriate, for diverse uses. It also must be flexible enough to respond to the continuous and rapid changes in scientific research and technology. While there are many technical issues to be resolved to take full advantage of past, current, and future investments in ICT infrastructure, the main barriers to effective data access and sharing are no longer technical, but are institutional and managerial, financial and budgetary, legal and policy, and cultural and behavioural.

6.1.1 Data Preparation and Metadata: ICPSR

In 1995, the Inter-University Consortium for Political and Social Research (ICPSR) initiated the development of the Data Documentation Initiative (DDI), an international criterion and methodology for the content, presentation, transport, and preservation of metadata about datasets in the social and behavioural sciences. DDI, which is in XML format, helps enhance users' ability to acquire and use data while it assists producers in packaging and disseminating them. After a period of beta-testing with participating international organisations, DDI is now in use by a number of organisations, including Networked Social Science Tools and Resources (NESSTAR), Health Canada, and ICPSR. ICPSR continues to assist data producers in preparing their data through its "Guide to Social Science Data Preparation and Archiving," a guide with broad appeal for individuals and organisations searching for easy and effective ways to technically manage and prepare data so that they can be easily and effectively placed into network environments (For more information on ICPSR and DDI, see Data Documentation Initiative (n.d.) and ICPSR (n.d.). For information on the importance and development of DDI, see Norwegian Social Science Data Services (1999)).

Technical operating principles for data access regimes include interoperability (of protocol and software to ensure the access and usability and multiple use of the data); and quality (including technical components of authenticity, integrity, and security) of data.

6.2 Institutional and managerial domain

While the core open access principle applies to all science communities, the diversity of the scientific enterprise suggests that a variety of institutional models and tailored data management approaches are most effective in meeting the needs of researchers.

Because scientific data have many different characteristics and uses, there is no monolithic institutional and management approach that can be applied universally. Key characteristics of data production and use include whether the data are (1) government-generated or generated at a research institution using public funds; (2) useful

only within the discipline or across many disciplines; (3) useful over the very long term or only within short-term horizons; (4) have public-policy implications; or, (5) have significant broader economic and social value, among other factors.

Institutional and managerial operating principles for data access regimes include transparency (systematic visibility of the data source); responsibility (explicit formal institutional rules on data management); and accountability (rendering public account for the performance of data access regimes).

6.2.1 Negotiated collaborations: CERN

The European Organisation for Nuclear Research, CERN, is one of the world's largest scientific laboratories, presently financed by twenty European countries. CERN overtly subscribes to the core principle and premises outlined in this report. However, the raw experimental data set does in itself not make much sense outside of the context of the specific experiment. The sheer size also necessitates heavy processing. Experiments at CERN are typically run by large-scale collaborations. Within each collaboration access to the data is unproblematic. In this stage the data are protected, however, partly because of technical issues (size and interpretation) and partly because of the competition between researchers. Availability of this type of data to other researchers depends on negotiations. At a higher level of interpretation CERN puts its data in the public domain. The cleaned up and interpreted data are made available to the international physics community in the form of a Data Summary Table. The type of data produced and the method of processing used will therefore play a large part in deciding upon the most effective management model to adopt. This flexibility of management approach is a key factor in the data production and sharing environment at CERN.

6.3 Financial and budgetary domain

Scientific data infrastructure requires continued, and dedicated, budgetary planning and appropriate financial support. The use of research data cannot be maximized if access, management, and preservation costs are an add-on or after-thought in research projects.

In many areas of public research, there are indications of discrepancies between the funding of the specific research itself and the related data-management requirements (which do not necessarily benefit the individual scientist, but which are necessary for data re-use). Generally, research organizations fund the former well, but pay scant attention to the latter. In the digital environment, scientific data sets must be viewed as a key element of the broader research infrastructure and as an investment in the future capacity to innovate and solve pressing problems. Adequate support is essential for data-management functions, such as the development of sufficient explanatory documentation (i.e., metadata) for each data set, conversion of old formats onto new media, adaptation to new standards, and long-term preservation, archiving, and maintenance.

Budgetary operating principles for data access regimes include operational efficiency (maximizing the return on investment by promoting re-use of data, and providing proper documentation, specialists, and effective data management facilities).

6.3.1 Funding schemes “on a rolling basis:” the European Bioinformatics Institute (EBI)

The official mission of the EBI is to ensure that the growing body of information from molecular biology and genomic research is placed in the public domain and is freely accessible to the scientific community in ways that promote scientific progress. Like other scientific bodies, the EBI has a major problem in the funding for its building, maintaining and making available databases and information services even though they represent only a small fraction of the total research costs. The key issue is that funding for data sharing infrastructures needs to be constructed “on a rolling” or on-going basis to maintain effective data management. These funding requirements are very different from the funding schedules of research, which are usually project oriented. These differences in budgeting constitute the main threat to the EBI's commitment to maintaining the public availability of its data.

6.4 Legal and policy domain

National laws and international agreements directly affect data access and sharing practices, despite the fact that they are often adopted without due consideration of the impact on the sharing of publicly funded research data.

Intellectual property laws, information policies, institutional guidelines, and contracts at the national and international levels often impose terms and conditions on data access and sharing practices. Laws and policies governing data access and sharing practices may vary among different countries, resulting in barriers to scientific cooperation and progress. Based on a recent Web survey (Wouters, 2002), most of the national research organization managers who responded expected that data sharing will become a major policy issue in the next five years. This situation requires greater attention by the science policy community at all levels. In particular, restrictions on re-use of public data by the research community must be eliminated or minimised as much as possible. Research grant provisions and licensing templates for promoting open access and unrestricted re-use of public research data already exist, but have not yet been broadly adopted.

Legal and policy operating principles for data access regimes include property (balance intellectual property rights of investigator and institution versus public good); and legality (lawful data management, respecting national security, privacy and trade secrets).

6.4.1 Policy interconnections: functional MRI and the Institutional Review Boards

The functional Magnetic Resonance Imaging Data Center's (fMRIDC) principal endeavour is to promote data sharing in brain mapping. The Western tradition of informed consent in bio-medicine operates according to the principle that the 'most specific consent is the best consent.' When data are to be gathered for submission to databases, the specificity of consent may run counter to the goals of meta-analysis or re-analysis by third parties, to investigate issues different from those for which the data was originally gathered. The creation of infrastructures for data sharing, therefore, has to conform to the rules of regulatory bodies, such as institutional review boards (IRBs), whose approval must be obtained to share data. As such, these bodies function as gatekeepers to the circulation of data. International coordination may also be necessary. Researchers submitting or requesting data across national boundaries may find it especially difficult to act in accordance with the various ethical guidelines that exist in different countries. The fMRIDC has been hesitant to accept data from non-US settings because of concerns regarding IRB compliance.

6.5 Cultural and behavioural domain

Appropriate reward structures are a necessary component for promoting data access and sharing practices. These apply to both those who produce and those who manage research data.

Although formal policy frameworks and regulations are necessary to make research data publicly available, they need to be supplemented by appropriate community-based norms and incentives for researchers to share and provide access to their data and for appropriate recognition of their data-related work. In many cases, there is a general lack of reward structures and mechanisms to promote open access to, and sharing of, data from public research.

Cultural and behavioural operating principles for data access regimes include quality (trust that data are what they purport to be); professionalism (build on codes of conduct and ethics of the scientific community); flexibility (there is no single model on how data access must be provided.)

6.5.1 Incentives: the Protein Data Bank

To publish in scientific journals, U.S. scientists involved in the field of crystallography must deposit their data in the Protein Data Bank (PDB) and acquire an accession number. As PDB Director Helen Berman explains, "By requiring everyone to submit data, the community is assured of having the most up to date information possible. Now, increasingly, under our regime, a lot of [data] depositors have come to realize that the practice that we use has some advantages for them in that we check things and we find errors and inconsistencies. That actually improves the quality of the product they produce."

7 POSSIBLE FOLLOW-UP STEPS FOR INTER-GOVERNMENTAL AND GOVERNMENTAL BODIES

Our findings from the case studies and from other research indicate a number of action areas by the different actors involved in making possible open access to, and sharing of, publicly funded research data. In this section we recommend possible action areas for the OECD and national governments.

7.1 OECD

As an international organization with credibility and stature in the science policy arena, the OECD, through the CSTP, can play a crucial role in promoting access to, and sharing of, data from publicly funded research. Central to this role is the gathering and sharing of information on successful practices in data related activities and policies. At the international level, only a few organizations have undertaken to do this, usually in the context of a specific discipline or research program. The recent, and vast, expansion of research data assets and the trend towards issue-based, interdisciplinary research, however, suggests that all countries and all fields of science stand to benefit from greater attention and an organized and coordinated approach to effective policy actions.

In its report to OECD, the Follow-up Group concluded:

The OECD should put the issues of data access and sharing on the agenda of the next Ministerial meeting. ICT advances have created the ability to transform science. New tools allow researchers to find data in seconds that would have taken months just a few years ago. Effective data access and sharing requires a comprehensive policy approach for implementation by public research institutions. Monitoring progress and devoting attention to the public research data issues and activities would assist decision-makers and research support agencies in developing appropriate policies and allocating resources.

This recommendation was made in March 2003. At the meeting of the OECD Committee for Scientific and Technological Policy (CSTP) at Ministerial Level on 30 January 2004 ministers responsible for national science and technology policies of OECD countries endorsed the *Declaration on Access to Research Data from Public Funding* (2004). In the Declaration, CSTP was invited to formulate OECD guidelines for *Access to Research Data from Public Funding*. A CSTP Working Group has been installed to draft these international guidelines.

7.2 National Governments

Although the OECD, UNESCO, ICSU, CODATA, and other international bodies can play a role in improving the current situation regarding research data access and sharing, it is at the national level that many important decisions and actions must be taken. National governments provide the resources for making data accessible, establish the overall policies for data management, regulate matters such as confidentiality and privacy, and determine restrictions based on national security. Most important, it is national governments that are responsible for the major research support and funding organizations, and it is here that many of the managerial aspects of data sharing need to be addressed.

The national governments of OECD countries should consider:

- 1. Adopting, and effectively implementing, the principle that data produced from publicly funded research should be openly available to the maximum extent possible.** The public investments made in research data collection can only be maximized if the data are preserved, managed, and made accessible. This requires coordinated attention by governments at all levels, and adequate policy and financial support. The starting point for these actions, however, is the affirmation that data collected using public funds should be openly accessible to all.
- 2. Encouraging their research funding agencies and major data producing departments to work together to find ways to enhance access to statistical data, such as census materials and surveys.** Many countries have taken steps to facilitate access to census and survey materials by developing catalogues, user-friendly repositories, off-site research facilities, training programs, and regulatory frameworks for providing appropriately guarded access to confidential information. Such steps have proven

enormously effective in maximizing the use of national surveys and producing insights into the functions of economies and societies.

3. **Adopting free access, or marginal cost pricing, policies for the dissemination of research-useful data produced by government departments and agencies.** The use of information collected through public funding should be freely accessible for research purposes. This maximizes the use of such information for public policy and public knowledge development.
4. **Analyzing, assessing and monitoring policies, programs, and management practices related to data access and sharing policies within their national research and research funding organizations.** This information would be useful to national governments so that they may assess the implementation of the previous three considerations. The resources, support programs, policies, and regulations related to research data sharing are, in large part, developed and implemented by research funding organizations. The operations of these organizations play a crucial role in determining the degree to which data are made accessible and shared between researchers. Many organizations, such as NSF and NIH in the United States, the Social Sciences and Humanities Research Council in Canada, and the European Science Foundation are now developing, or have developed, policies, regulations and support programs that promote data sharing. Issues such as establishing protocols for the collection and release of confidential information, developing technical infrastructure, agreeing on metadata standards, requiring data preservation strategies within individual research projects, and including data management costs as eligible expenditures in grant applications have been dealt with by one or more of these agencies. It would benefit the global scientific community if decision-makers within national governments had a clear understanding of where their respective agencies stood in relation to those in other countries.

7.3 Areas for Further Examination

The OECD/CSTP Working Group currently engaged in drafting international guidelines will consider the other recommendations to OECD from the report of the Follow-up Group (OECD Follow-up Group on Issues of Access to Publicly Funded research Data, n.d.). The recommendations concern the following activities:

1. **Consider conducting or coordinating a study to survey national laws and policies that affect data access and sharing practices.** This relatively simple undertaking could determine what policies exist, how accessible they are, and result in listing of the web sites where these policies are posted. This study would be of considerable benefit to science policy-makers, research administrators, and information resource managers in all countries, both within OECD and beyond. The study could look at the feasibility of developing a central and easily accessible repository of national laws and policies that affect data access and sharing practices. Such a compilation does not currently exist, and could be useful to facilitate international research collaborations (for a preliminary survey, see Wouters, 2002).
2. **Consider conducting or coordinating a study to compile model licensing agreements and templates for access to and sharing of publicly funded data.** Depending on the context, numerous factors need to be considered in data access and sharing arrangements. Nevertheless, many contractual models already exist that have been developed by research funding organisations, research program managers, university administrators, librarians, and others. The OECD, as a global organization, is ideally suited to span national domains where examples do exist, and thereby bring an international perspective. The study could compile and review existing agreements and models to find exemplary approaches. Having readily available models on hand would be of considerable benefit to researchers, universities, and research institutions, as well as data centers and archives, and could facilitate international research collaboration.
3. **Governments from OECD countries should consider expanding their policy framework of research data access and sharing to include data produced from a mixture of public and private funds.** Collaborative public/private research projects, and the resulting data, have their own unique set of characteristics and issues. As more national governments promote public-private partnerships in research, these issues will be of increasing importance to both public researchers and the companies that are involved. A further examination of the state of data sharing and access in these types of research arrangements needs to be made to develop sound science policy guidance.
4. **Consider examinations of research data access and sharing to include issues of interacting with developing countries.** The increase of participation in the research enterprise benefits the global science system and innovation. Providing developing countries with access to data from publicly funded research increases their participation in science. Further, as United Nations Education, Scientific and Cultural

Organization (UNESCO), the International Council of Scientific Unions (ICSU), private foundations, and other organizations have emphasized, access to scientific knowledge by developing countries is vital to the progress of the entire world. This access is particularly important in the context of global issues such as population health, environmental change, and food production. Of course, open access to data from publicly funded research in developed countries can provide a valuable resource for economic development, education, and scientific capacity building. Many efforts are already underway to improve access for researchers in developing countries (e.g., providing free or below-costs access to data and scientific information), as well as establishing optimal data regimes for developing countries to share their data (e.g., addressing issues of data repatriation). A systematic examination of barriers and best practices would provide both a picture of the current situation and a set of guidelines for further action.

- 5. Consider promoting further research, including a comprehensive economic analysis of existing data access regimes, at both the national and research project or program levels.** To date, no one has yet undertaken a comprehensive, economic analysis of different data access regimes. Several key issues have not been closely examined, including the relative costs of providing data openly or not, the impact of cost recovery on the use of those data, and the positive externalities and network effects from providing open access to publicly funded research data. The OECD should consider conducting this type of analysis or encouraging member country research organizations to fund such studies.

8 CONCLUSION

Improving access to and sharing of publicly funded research data is an issue that touches on all aspects of the research enterprise and the development of knowledge, and involves all participants in the conduct of research. For the individual researcher, the sharing of data, particularly prior to publication, can be burdensome, time consuming, and unrewarding if the necessary measures are not taken to provide funding, facilities, and a social context that emphasises its value to the research community and to society (Sacrifice for the greater good?, 2003).

Advances in ICTs, the internationalisation of science, and the trend toward issue-based research hold great potential for the advancement of knowledge and for the benefit of all people. This potential will not be fully realized unless all of the major elements of data access regimes identified in this report are properly developed. To do so will take considerable discussion, understanding, and commitment on the part of all those involved in research, particularly at the policy and funding levels.

Agreement among OECD governments on a set of general principles to shape specific data access regimes, as well as adoption of the recommendations set forth above, would be enabling for scientists, empowering for citizens, and provide an important contribution to fulfill the promises of e-science.

9 ACKNOWLEDGMENTS

This article is a slightly updated version of the report of the same name (DSTI/STP(2003)20 dated 19 March 2003 and submitted to the Committee for Science, Technology and Industry (CSTP) of the Organization for Economic Cooperation and Development (OECD) at its March 2003 meeting by the *OECD Follow-up Group on Issues of Access to Research Data from Public Funding to the OECD*.

The Follow-up Group was established shortly after the March 2001 OECD Committee on Science and Technological Policy (CSTP) meeting, and delivered its report to the CSTP in March 2003. The March 2003 OECD Report Appendix of case studies, not included here, has been expanded and published elsewhere (Wouters & Schröder, 2003). The term "Follow-up" refers to recommendations that came from the Global Research Village III Conference on Access to Publicly Financed Research (December 2000, Amsterdam), to be followed-up by the fourth Global Research Village Meeting held in 2002 in Warsaw.

The recommendations of the report led to the establishment of a CSTP Drafting Group that prepared the Declaration on Access to Research Data from Public Funding (2004) for the ministerial CSTP meeting and subsequently (in 2004) to the establishment of the CSTP Working Group that is currently preparing the OECD guidelines on the subject. The issues raised by the recommendations of the report were addressed by the Drafting Group and will be further explored by the current Working group.

On the basis of this March 2003 Report (DSTI/STP(2003)20), a CSTP Drafting Group prepared the documents that led to the *Declaration on Access to Research Data from Public Funding* endorsed by ministers at the meeting of the OECD Committee for Science and Technological Policy at Ministerial Level, Paris, 29 to 30 January 2004, Final Communiqué of *Science, Technology, and Innovation for the 21st Century*, Annex 1, on 30 January 2004. In the Declaration, OECD Ministers invited CSTP to formulate OECD guidelines for *Access to Research Data from Public Funding*. A CSTP Working Group has been installed to draft these international guidelines.

The authors would like to acknowledge participants in the members of the OECD Follow-up Group Sigrun Eckelmann, Deutsche Forschungsgemeinschaft (DFG) – Germany; Tim Hubbard, Wellcome Trust Sanger Institute – UK; Koji Kamitani, MEXT (Ministry of Education, Culture, Sports, Science & Technology) – Japan; Doug McEachern, Australian Research Council – Australia; Masamitsu Negishi, National Institute of Informatics – Japan; Andrzej P. Wierzbicki, National Institute of Telecommunications – Poland; Jan Windmueller, Ministry of Science, Technology and Innovation – Denmark. The following people contributed their expertise to the project: Jacky Bax, Ministry of Education, Culture and Science - The Netherlands; Kathleen Cass, CODATA - Paris, France; Gudrun Maas, OECD, Directorate for Science, Technology and Industry Science and Technology Policy Division – France; Tony Mayer, European Science Foundation – France; David Schindel, National Science Foundation – France; Hugo von Linstow, Global Biodiversity Information Facility – Denmark; Colin Reddy, Networked Research and Digital Information (Nerdi) - The Netherlands; and, Teri Simas, University of California, San Diego – USA.

Participants from the United States were supported by the NSF through grant ACI-9619020; co-funded by the Office of International Science and Engineering, the Divisions of Shared Cyberinfrastructure, Biological Infrastructure, and Social and Economic Sciences and the Directorate for Math and Physical Sciences. The Netherlands' Ministry of Education, Culture and Science sponsored the two studies on data access published by NIWI-KNAW. Other support was provided by the Secretariat of the international CODATA (Committee on Data for Science and Technology), Polish State Committee for Scientific Research, and the European Science Foundation. The views expressed by the authors are their own and not necessarily those of their employers or supporting agencies.

10 REFERENCES

Access all areas (2004) *The Economist*. 372 (8387), 73.

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., Moorman, D., Uhler, P., & Wouters, P. (2004) An international framework to promote access to data. *Science* 303, 1777-78.

Atkins, D. E., Droegemeier, K.K., Feldman, S.I., Garcia-Molina, H., Klein, M.L., Messerschmitt, D.G., Messina, P., Ostriker, J.P., & Wright, M.H. (2003) *Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Arlington, VA: National Science Foundation.

Biomedical Informatics Research Network (n.d.) Homepage of Biomedical Informatics Research Network. Available from: <http://www.nbirm.net>

Colwell, R. (2002) A Global Thirst for Safe Water: The Case of Cholera. *Abel Wolman Lecture at the National Academy of Sciences*, Washington, D.C. Retrieved February 2, 2003 from the World Wide Web: http://www7.nationalacademies.org/wstb/2002_Wolman_Lecture.pdf

Cook, A. (2004) Open access to the US government urged. *The Scientist* (July 21). Retrieved July 21, 2004 from the World Wide Web: <http://www.biomedcentral.com/news/20040721/01/>

Data Documentation Initiative (n.d.) About the Data Documentation Initiative Alliance. Retrieved from the *Data Documentation Initiative* website: <http://www.icpsr.umich.edu/DDI/ORG/index.html>

Declaration on access to public research data from public funding (2004) Annex I to Final Communiqué of OECD Committee for Science and Technological Policy at Ministerial Level: *Science, Technology, and Innovation for the 21st Century*, Paris, 29 to 30 January 2004, Final Communiqué, (www.oecd.org).

Esanu, J. M., Uhlir, P. F. (Eds.) (2004) *Open Access and Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium*, Washington, D.C.: National Academies Press.

European Union Green Paper (1998) Public Sector Information: A Key Resource for Europe, COM (98) 585. Retrieved March 3, 2002 from the World Wide Web: [europa.eu.int/ISPO/docs/policy/docs/COM\(98\)585/](http://europa.eu.int/ISPO/docs/policy/docs/COM(98)585/)

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds) (1985) *Sharing Research Data*, Washington, D.C.: National Academy Press.

Franken, H. (2001) Conference Conclusions. In Schröder, P. (Ed.), *Access to Publicly Financed Research*, Amsterdam, NIWI, 75-82. Retrieved August 10, 2004 from the Web: <http://www.esf.org/sciencepolicy/172/report.pdf>

Global Biodiversity Information Facility (n.d.) Homepage of the Global Biodiversity Information Facility. Available from: <http://www.gbif.net>

House of Commons Science and Technology Committee (2004) Scientific Publications: Free for all? 10th report of session 2003-2004. Retrieved January 26, 2004 from the World Wide Web: <http://www.publications.parliament.uk/pa/cm/cmsctech.htm>

ICPSR (n.d.) Guide to social science data preparation and archiving. Retrieved from the ICPSR website: www.icpsr.umich.edu/ACCESS/dpm.html

Kaul, I., Grunberg, I., & Stern, M. (Eds.) (1999) Defining Global Public Goods. *Global Public Goods: International Cooperation in the 21st Century*, New York: Oxford University Press.

Medical Research Council (n.d.) Draft MRC statement on data sharing and preservation policy. Retrieved July 19, 2004 from the *Medical Research Council* website: http://www.mrc.ac.uk/index/strategy/strategy-science_strategy/strategy-strategic_implementation/strategy-data_sharing/strategy-data_sharing_policy-link

McMahon, R.C. (1996) Cost Recovery and Statistics Canada. *Government Information in Canada* 2(4). Retrieved February 10, 2003 from the World Wide Web: <http://www.usask.ca/library/gic/v2n4/mcmahon/mcmahon.html>

Mervis, J. & Stokstad, E. (2002) NASA censors report on agriculture threats. *Science* 297, 1973-75.

NIH Office of Extra Mural Research (n.d.) NIH Draft Statement on Sharing Research Data. Retrieved from the NIH website: http://grants2.nih.gov/grants/policy/data_sharing/

National Research Council (1997) *Bits of Power: Issues in Global Access to Scientific Data*, Washington, D.C.: National Academy Press.

National Research Council (1999) *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*, Washington, D.C.: National Academy Press.

Norwegian Social Science Data Services (1999) Providing Global Access to Distributed Data through Metadata Standardisation -- The Parallel Stories of NESSATAR and DDI, submitted by the Norwegian Social Science Data Services to the Conference of European Statisticians, UN/ECE Work Session on Statistical Metadata, Geneva, Switzerland, 22-24 September 1999. Retrieved from the World Wide Web: <http://www.nesstar.org/papers/GlobalAccess.html>.

OECD Committee for Science and Technological Policy (2004) *Science, Technology, and Innovation for the 21st Century. Meeting of the OECD Committee for Science and Technological Policy at Ministerial Level, 29 - 30 January 2004 - Final Communiqué, Annex 1*. Retrieved July 19, 2004 from the OECD website: http://www.oecd.org/document/15/0,2340,en_2649_34487_25998799_1_1_1_1,00.html.

OECD Follow-up Group on Issues of Access to Publicly Funded research Data (n.d.) The public domain of digital research data. Retrieved July 19, 2004 from the World Wide Web: <http://dataaccess.ucsd.edu>

Pira International (2000) Commercial exploitation of Europe's public sector information: Final report for the European Commission Directorate General for the Information Society, UK, 30 October 2000. Retrieved from the web January 20, 2004 from the World Wide Web: http://europa.eu.int/information_society/topics/multi/psi/docs/pdfs/commercial_exploitation/commercial_final_report.pdf

Research Councils UK (n.d.) About the UK e-Science programme. Retrieved July 19, 2004 from the World Wide Web: www.research-councils.ac.uk/escience/

Sacrifice for the greater good? (2003) *Nature* 421, 875.

Stiglitz, J., Orszag, P. & Orszag, J. (2000) *The Role of Government in a Digital Age*. Retrieved March 2, 2002 from the World Wide Web: <http://www.sbgo.com/Papers/CCIA%20FINAL.pdf>.

Stix, G. (2001) Triumph of Light. *Scientific American* January, 80-85.

Suber, P. (2004) UK House Committee releases its report on open access. *Open Access News*. Retrieved July 19, 2004 from: http://www.earlham.edu/~7Epeters/fos/2004_07_18_fosblogarchive.html#a109028453862279542

Weiss, P. (2002) *Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impact*, Washington, D.C.: U.S. National Weather Service. Retrieved February, 5 2003 from the World Wide Web: www.weather.gov/sp/Borders_report.pdf.

Weiss, P. (2003) Conflicting International Public Sector Information Policies and their Effects on the Public. In Esanu, J. M. & Uhlir, P. F. (Eds.), *The Role of Scientific and Technical Data in the Public Domain: Proceedings of a Symposium*, Washington, D.C.: National Academies Press, 29-132.

Westbrook, J. H. & Grattidge, W. (1991) A Glossary of Terms Related to Data, Data Capture, Data Manipulation, and Databases. *CODATA* 23 (1, 2).

Wouters, P. (2002) *Policies on Digital Research Data: an International Survey*. Amsterdam, Netherlands Institute for Scientific Information Services, NIWI-KNAW.

Wouters, P. & Schröder, P. (Eds.) (2003) *Promise and practice in data sharing*. Amsterdam, Netherlands Institute for Scientific Information Services, NIWI-KNAW.

Bonner zoologische Beiträge	Band 51 (2002)	Heft 3/4	Seiten 205-212	Bonn, August 2003
-----------------------------	----------------	----------	----------------	-------------------

Computerizing Bird Collections and Sharing Collection Data Openly: Why Bother?

A. Townsend PETERSON & Adolfo G. NAVARRO-SIGÜENZA

Natural History Museum, Lawrence, Kansas, USA and Museo de Zoología, México, México

Abstract. Natural history museum collections provide the basic documentation of life on Earth. As such, they represent the critical and unique resource by which that life may be understood, and have immense economic and scientific importance. Nevertheless, particularly in recent decades, natural history museums have seen less and less attention – and resources – in spite of their importance. A series of new efforts, however, aim to recoup that prominence via community efforts to unite data resources towards a vastly improved understanding of biodiversity and its implications. The Species Analyst represents an effort to unite natural history collections databases worldwide to this end: 77 institutions now cooperate or are committed to cooperate in serving records of 51 million natural history museum specimens to users worldwide, and has seen more than 700,000 users to date.

Key words. The Species Analyst, TSA, archive for biodiversity, worldwide facilities

1. INTRODUCTION

Computerization of ornithological collections is increasingly considered a priority for curators and staff of natural history museums. A common quandary, however, is how and why to get started. The curator is presented with a bewildering variety of databasing programs, some especially designed for specimen records, and others off-the-shelf generic database programs that can be customized for any use. Choice of a platform, choice of data fields, and choice of computerization strategy all become critical – and difficult – consideration. Unfortunately, these considerations can often seem so complex that computerization efforts are not initiated.

Moreover, presented with a thousand and one other priorities of collections building, specimen conservation, institutional politics, and research efforts, and given the significant time investment that computerization requires, the question arises as to whether the result is worth the time. That is, one must consider what are the benefits of computerization, and how much do they benefit the collection, the curator, and the broader community.

The purpose of this contribution is to provide a rationale for computerizing bird collections as a critical step forward in their care. Along the way, we review steps involved – a sort of minimum-standard guide to starting computerization efforts. Finally, we provide a series of examples of how computerizing collections data, and sharing those data across many institutions worldwide, benefits the collections themselves.

2. WHY COMPUTERIZE A COLLECTION?

Databasing or computerizing a collection is a lot of work, and may easily absorb years of effort. So why do it? Several reasons argue strongly for taking this step. A partial list follows:

- *Get to know your collection* – a sweep through the whole collection, drawer by drawer, gives a unique knowledge of a particular collection.
- *Discover important specimens* – many fascinating discoveries have resulted from the specimen-by-specimen attention during computerization efforts, including species new to science, lost type specimens, important historical specimens, etc.
- *Detect problems* – again, the specimen-by-specimen attention can help to detect serious problems that might otherwise not be noticed ... damage from insects or water, fading of plumages, drying of spirit specimens, etc.
- *New views of the collection* – although we are familiar with summaries of collections in terms of taxonomic completeness, and perhaps regional summaries, many new views of collections open when a collection is computerized, e.g., maps of the geographic distribution of specimens, summaries of accessions over time, etc.
- *Save curatorial time* – making summaries of holdings, preparing loan invoices, tracking down particular specimens, and many other curatorial tasks are considerably more efficient when the collection is available in database form.
- *Standardize taxonomy* – once data are in electronic form, comparing names against a standard list (e.g., the Peters' check-list) can identify a first set of non-

- standard names that require checking and updating.
- *Efficient information access* – many questions and data requests that require hours or days of work for an uncomputerized collection will suddenly become feasible to answer in minutes, making possible much more creative uses of the information in collections. For example,
 - What are your holdings of taxon X?
 - What are your holdings from country X?
 - Do you have specimens collected by person X?
 - What is the history of specimen acquisition rates in your collection?
 - And many more ...

In short, computerization of a collection is a major undertaking, but ends up repaying the investment of time and effort many times over.

3. CHOOSING A PLATFORM

The first big question to be answered is about which platform (databasing program) to use. This decision becomes complex ... sometimes, museum administrators decide to force all collections in the museum to use the same program. Even if one has the freedom to choose, should one choose among the many programs that have been developed specifically for natural history museum specimens (BIOTA, BIOTICA, SPECIFY, etc.), or a generic program off the shelf (e.g., Microsoft Access, Ora-

cle)? Regarding this choice, each option has its strengths and weaknesses (Table 1). In general, we would recommend the off-the-shelf option for small, old, or inactive collections, and the specimen databasing programs for larger, data-rich, and very active collections.

Regardless of this choice, one should insist on several minimum criteria for a databasing platform. These criteria are critical features of a program that must be fulfilled in order to avoid problems. As follows:

- Capacity for export to other, generic formats, particularly ASCII delimited format, to allow reporting, export to other programs, and porting to future technologies and platforms.
- Compatible with Standardized Query Language (SQL), which permits many functionalities to be added to your database related to sharing data.

Once a platform has been identified that fits the particular needs of a collection, and meets these basic requirements, then design of the computerization effort can begin.

If the reasoning outlined above suggests that the best solution to computerization is that of a more complex program specifically designed for natural history specimen data, then you should read about several of the programs that are available. Links to a number of such programs are presented in Table 2.

Table 1: Summary of advantages and disadvantages of specialized versus generic programs as platforms for computerizing bird collections.

NATURAL HISTORY MUSEUM SPECIMEN DATABASING PROGRAMS	OFF-THE-SHELF GENERIC DATABASING PROGRAMS
<p>Advantages</p> <p>Designed specifically for specimen management</p> <p>Features such as authority lists, loan invoice reporting, etc.</p> <p>No customization or little customization required</p> <p>Most complex solutions specific to natural history specimens are tractable</p>	<p>Long-term continuity of support from the company</p> <p>Easy availability of expert advice, given broad usage in many communities</p> <p>Simplest solutions are feasible</p> <p>Simple learning curve</p>
<p>Disadvantages</p> <p>Can disappear – long-term support often depends on a person – researcher or developer – who can decide not to support the program further, or who may decide not to update to newer versions (e.g., MUSE)</p> <p>Expert advice may be unavailable in a particular city</p> <p>May not permit very simple solutions to simple problems</p> <p>Steeper learning curve</p>	<p>May need customization of program for intermediate-to-complex situations</p> <p>Not designed specifically for specimen data</p> <p>Complex features (e.g., reporting, authority lists) not automatically available</p>

Table 2: Selected specialized programs designed specifically for collections data. Provided are World Wide Web links for more information.

Program	URL
SPECIFY	http://usobi.org/specify/
Biótica	http://www.conabio.gob.mx/biotica_ingles/distribucion_b.html
BioLink	http://www.biolink.csiro.au/
BIOTA	http://viceroy.eeb.uconn.edu/biota
KE EMu (not recommended for integration via <i>Species Analyst</i>)	http://www.kesoftware.com/

4. CHOOSING DATA FIELDS

This step may prove to be the most critical of all in the process of computerization. With too many fields, time and filespace are wasted, whereas with too few, they will have to be added later or one will have to live without them. If an incorrect structure is chosen, the database may be forever handicapped by this design flaw. However, the challenge is reduced quite a bit with an understanding of a few basic ideas. Specimen data, in their simplest form, distill down to three linked sets of information about each specimen:

- *Taxonomic information* – the taxonomic identity of the specimen
- *Geographic information* – the geographic location of its collection
- *Detailed documentation of the specimen* – time of collection, collector identity, museum catalogue number, sex, age, body mass, etc.

Thinking in this manner, we can envision a structure for a specimen database that would capture this information optimally. Taxonomy and geography are both hierarchical concepts, and so we can represent them as such, which would make for three interacting sets of information (Fig. 1).

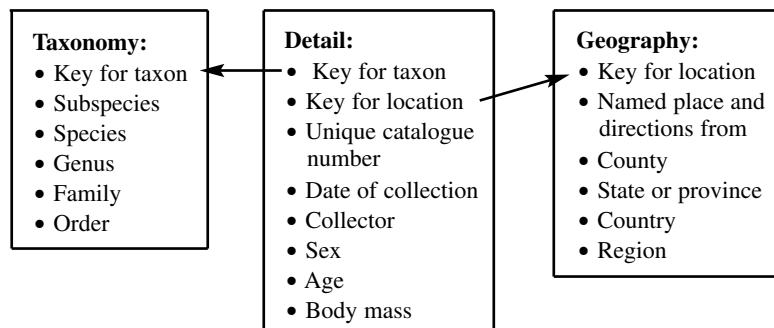


Fig. 1: Diagrammatic illustration of a simple relational database structure designed to link hierarchically organized geographic and taxonomic information with specific data regarding a particular specimen.

In the simplest sense, then, even in a spreadsheet program such as Microsoft Excel, or (better still) as a single table in a database program such as Microsoft Access, one could use a straightforward single table that holds critical fields (see Table 3). This very simple structure provides a clear, workable solution for small collections. In a more complex situation, in which more specimens are to be computerized, this structure can be made relational (Fig. 1) (that is, made up of several tables that interconnect). The advantage of a relational database structure is that elements of the database are entered only once: e.g., the locality descriptor for the 150 specimens collected at USA/Kansas/Douglas Co./Lawrence/10 km E is entered only once, reducing the possibility of typographical errors.

Table 3: Critical minimum set of fields for a simple collections database.

Field	Example
Catalogue number	15230
Genus	Cyanocitta
Species	cristata
Subspecies	cristata
Date of collection	24 October 1956
Collector	Fredrick E. Jones
Sex	Female
Age	Adult
Body mass	120 g
Country	USA
State or province	Ohio
County	Butler Co.
Named place and directions from	Oxford, 10 km E

This sort of simple relational structure can be implemented in a program such as Microsoft Access with a few hours' attention by a technician familiar with the program. The custom specimen database programs use a more complex relational structure, but one that is in essence based on this overall backbone. Again, the more complex the demands that one will wish to place on the database (e.g., more complex queries, more detailed reporting, more specimens), the more complex the database structure that will be required. For relatively simple applications, however, the simple flat file (single table) setup described above will often be adequate.

5. COMPUTERIZATION STRATEGY

The next question to be faced is the strategy for computerization. This decision depends heavily on the exact situation of a collection. If, on the one hand, an excellent paper catalogue or card file exists, one may wish to computerize directly from that, and then verify the accuracy and completeness later from the actual specimens. If, on the other hand, a good card file or catalogue does not exist, or if many specimens may have been omitted (exchanged or deaccessioned) or not entered in the catalogue, then you may be better off computerizing directly from specimens.

In general, two passes through the collection will be necessary as part of any computerization effort. The first will simply get each specimen's data into the computer as efficiently as possible. The second will verify (1) the existence of the specimen, (2) that all data elements are entered in the database, and (3) that all of the specimen's data are correct as entered. This verification step, although labor-intensive, is critical to making the database a correct representation of the information contained in the specimens' labels.

All computerization efforts should involve the critical step of backing up data at regular intervals. Too many 'impossible accidents' have removed a year of work, and set a computerization effort back terribly. Backing up data should be done as permanently as possible ... that is, compact disks are better than floppy disks. It should also be done with redundancy: each time that you make a copy, if at all possible, it should not over-write the previous copy. This preservation of 'versions' of the database allows one to go back a week or a month if some error appears in the data set. Finally, given the possibility of more catastrophic losses, the back-up copies should be stored off-site, preferably in several places. Excellent storage sites for these copies can include libraries or archives, or curators' homes, or they can even be transferred via the Internet or via mail to another country.

6. THE SPECIES ANALYST (TSA)

The Species Analyst (<http://speciesanalyst.net/>) is a collection of software tools that permits integration of computerized collections data among institutions around the world into a distributed biodiversity information facility. For example, a user might wish to ask for records of any taxon from Yellowstone National Park or from Burma, or all specimens collected by Alexander von Humboldt, and retrieve information in a matter of seconds from 50 institutions around the world.

TSA uses a hybrid of Z39.50 (an information transfer protocol developed about 20 years ago in the bibliographic community) and XML (a more modern and efficient protocol) to permit efficient query and

retrieval of data. TSA may be accessed via a web portal that permits basic queries, or via extensions to Microsoft Excel (for retrieval of data in spreadsheet format) and ArcView (for retrieval of data as GIS coverages) (downloads available at <http://speciesanalyst.net/downloads>).

TSA currently integrates data sets from 22 institutions, for a total of 15 million specimen data records for over 50,000 species; a total of 58 institutions has committed to participation formally, which will take the total number of specimen records served to about 50 million. A special strength at present is in ichthyological data, as FishNet (<http://speciesanalyst.net/fishnet/>) has taken excellent advantage of TSA technology to create a data facility linking most important computerized fish collections. Now funded is a parallel network for mammal collections data (MANIS, based at the Museum of Vertebrate Zoology; <http://elib.cs.berkeley.edu/manis/>), and networks for herpetological and ornithological (expanded) specimen data are pending and in preparation, respectively.

7. WHY SHARE DATA ONCE COMPUTERIZED?

Above, we listed the first set of benefits of computerization of bird collections – namely, freer and more complete access to the information content of the specimens that make up the collection. These benefits are indeed considerable, and add enormously to a curator's ability to take care of a collection. However, once data are computerized, if they are shared, and integrated with data from other collections around the world, an additional set of benefits accrues.

In essence, a set of emergent properties comes into being once all (or nearly all) data are integrated for a particular taxon or region. We have come to appreciate these emergent properties as we have assembled the Atlas of Mexican Bird Distributions (NAVARRO & PETERSON, in prep.), a centralized database now including the contents of more than 60 natural history museum collections of Mexican birds. This 11-year project has resulted in a diversity of synthetic publications regarding the Mexican avifauna (NAVARRO-SIGUENZA et al. 1992a, b; PETERSON 1993; PETERSON et al. 1993; PETERSON 1998; PETERSON et al. 1998a, b; NAVARRO-SIGUENZA & PETERSON 1999, 2000; PETERSON et al. 2000, 2001, 2002). Herein, we will use this exemplar data set to demonstrate a variety of potential benefits to broad integration of data across institutions, as follows:

7.1 Georeferencing as a Community

Georeferencing locality data for specimens opens doors to a multitude of new capabilities and new func-

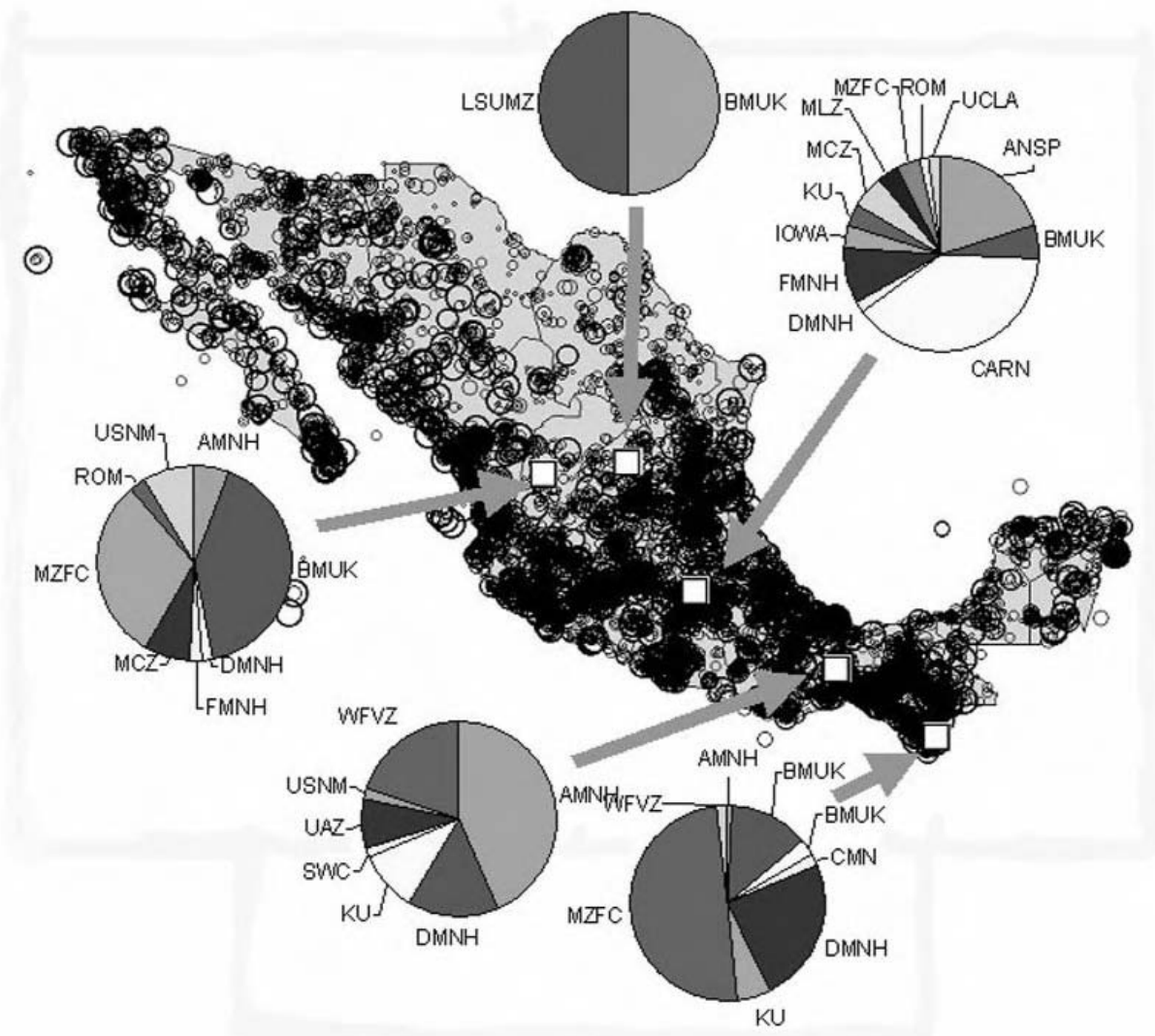


Fig. 2: Map of Mexico with collecting localities plotted by numbers of specimens collected at each point (graded symbol size: smallest = 1 specimen, largest = >100 specimens). For five points, to illustrate the redundancy of collecting localities among museums, we provide pie diagrams that illustrate the relative holdings of specimens from that particular site among scientific collections (see Acknowledgements for institutions and abbreviations).

localities to collections data. Indeed, all of the advances of geographic information systems (GIS) open up to collections data once latitude and longitude data are available for the collecting localities for each specimen. Nevertheless, georeferencing collections data – even once they are in electronic form – represents an enormous task.

Integrating this task over many institutions, however, takes advantage not just of having more people to help in a large task, but also of the redundant nature of the geographic sampling of birds (Fig. 2). Indeed, more than 25 % of Mexican bird collecting localities occur in more than one museum, and some in more than 20 museums. This redundancy results from col-

lections being dispersed among numerous museums (e.g., the specimens of Wilmot W. BROWN from Chilpancingo, Guerrero), and from certain sites being especially accessible or well-known as collecting localities in particular regions (e.g., Cerro San Felipe, Oaxaca).

A first experiment in cooperative georeferencing is beginning in the mammal community in the United States. The MANIS network, a U.S. National Science Foundation-funded effort, is connecting 17 institutions with computerized holdings of mammal specimens. A first step in MANIS' integration efforts is the pooling of institutional lists of localities to be georeferenced; institutions are then 'signing up' for particu-

lar regions, perhaps a home state, or an area of particular interest to the curator. In this way, efforts in georeferencing have a direct return for a particular investigator or institution, and add to the community pool of georeferenced information.

7.2 Detecting Errors in Date and Locality

Once specimen data are integrated, and have been georeferenced, further data refinements are possible. A common question is that of the relative reliability of the data associated with specimens from different collectors (BINFORD 1989). Because of the fragmented and dispersed nature of collector's material it has always been out of reach before. For instance, the still-living collector and ornithologist Robert W. DICKERMAN has deposited specimens at 14 of the 32 museums included in our present summary; the early twentieth century collector Wilmot W. BROWN has specimens distributed across 23 of the 32 museums. Once these data are pooled, however, new insights become possible regarding collectors' relative reliability.

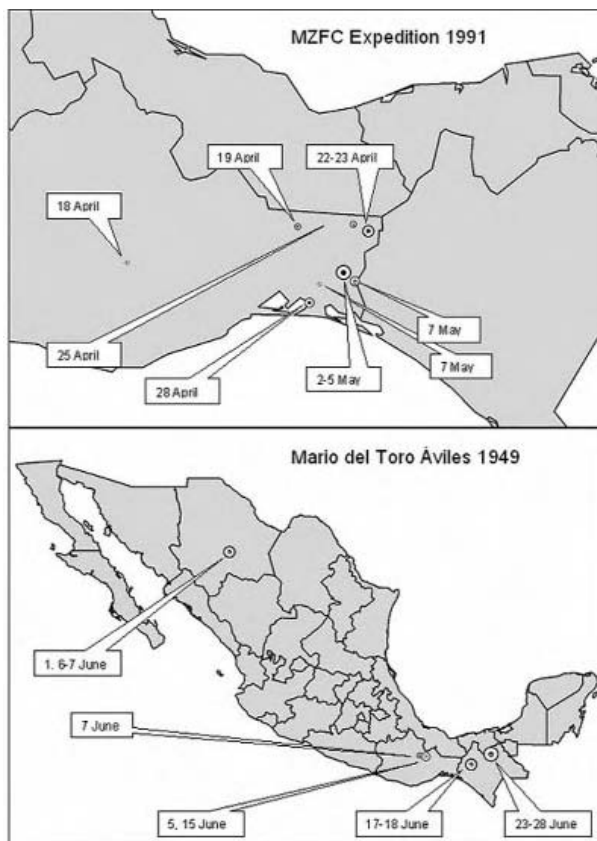


Fig. 3: Maps of collecting localities for two contrasting groups of collectors in Mexico: a Museo de Zoología, UNAM (MZFC) expedition in Spring 1991, and the collections of Mario del Toro Áviles in June 1949. Organized by collections date, consistencies and inconsistencies of specimen labeling become clear.

Basically, by assembling the entire opus of a collector, and sorting specimen locality by collecting date, it is possible to assess how geographically reasonable the combination of dates and localities is. Hence, to present a contrasting pair of examples, a Museo de Zoología, UNAM, expedition in 1991 scouted numerous sites in central and eastern Oaxaca (Fig. 3, top); although its route was complex, specimens from particular localities were clumped in time, and a sensible route could be reconstructed (although, in constructing this example, we detected an error in our georeferencing ... the 'Benito Juárez' referred to in the locality descriptor was the one in eastern Oaxaca, not the one in central Oaxaca). In stark contrast, specimens scattered across four museums (MLZ, LACM, FMNH, USNM) suggest that the infamous collector Mario del TORO ÁVILES worked at several sites across Mexico in June 1949; plotting these localities by date, however, reveals a number of points at which impossibly long journeys would have had to have been made in too short a time (Fig. 3, bottom). This result confirms earlier suspicions that del TORO ÁVILES' dates and localities are to be regarded with utmost caution (BINFORD 1989; PETERSON & NIETO-MONTES DE OCA 1996).

This approach can be used to detect problems in collectors' series, which will either be errors in date of collection or in collecting locality. Indeed, for an integrated, distributed data set consisting of the holdings of many institutions, it could be implemented as an error-seeking module that scans the data set collector by collector, and flags particular records as potential problems. These flagged specimen lists could then be distributed to collection curators for checking.

7.3 Detecting Errors in Identification or Georeferencing

A further refinement to specimen data also becomes possible, which will detect problems either in species identification or in georeferencing of localities. In essence, by viewing large quantities of occurrence data for a particular species, it is possible to detect spatial outliers, which likely represent identification or georeferencing problems. This process can be refined still further via ecological niche modeling for species: the ecological needs of a species are modeled (PETERSON 2001; PETERSON et al., in press) using high-end computational tools (STOCKWELL & NOBLE 1992; STOCKWELL 1999; STOCKWELL & PETERS 1999). These procedures use known occurrences of a species to produce a geographic view of areas meeting and not meeting its ecological needs; overlaying the same known occurrence points used to build the models allows identification of outlier occurrences.

As an example of this approach, we used the known occurrences of the brush-finch *Atlapetes pileatus* to



Fig. 4: Map of known collecting localities for the brush-finch *Atlapetes pileata*, overlain on a map of regions fitting the modeled ecological needs of the species (in gray), showing an old coastal locality in Tamaulipas as falling outside of the species' ecological niche.

build an ecological model and identify areas of appropriate and inappropriate ecological conditions for the species (Fig. 4). The modeling algorithm used is detailed elsewhere (STOCKWELL & NOBLE 1992; STOCKWELL 1999; STOCKWELL & PETERS 1999; PETERSON 2001; PETERSON et al., in press), but the result is that all known occurrence points fall into areas predicted to be appropriate for the species except one. This point (Fig. 4) represents an old locality on the coast of Tamaulipas, in the lowlands of eastern Mexico. The ecological modeling procedure identifies this site as a specimen locality that is not within the ecological possibilities of the species, and most likely represents an erroneous locality designation.

Like the collector itinerary approach, a procedure based on ecological niche modeling could be implemented as an error detection facility. A computer could periodically scan the pooled data resources for known occurrence points of each species, build ecological niche models for each species, and detect occurrence points that fall outside the ecological limits of the species. These points can then be flagged for checking by curators or collections staff.

7.4 Community-wide Activities: The Power of Numbers

Much more generally than for the preceding examples, it is important to emphasize the power of working of a community. When a proposal stems from a Division of Mammalogy at a particular museum, it carries far less force than a proposal that comes from all of the Mammalogy divisions from 17 institutions. This power of numbers – working as a community – makes possible many bold new funding initiatives.

Indeed, in the Species Analyst effort, several such community proposals have already been prepared, and have proven enormously successful. Proposals have been prepared and funded for a pilot North American bird network (U.S. National Science Foundation, funded 1998), a 15-member fish data network (U.S. National Science Foundation and U.S. Office of Naval Research, funded 2000), and a 17-member mammal data network (U.S. National Science Foundation, funded 2001). This success clearly results from the community nature of the proposals, and has resulted in more than \$2 million of new funding being available to the systematics collections community.

More generally, community efforts constitute an important step towards demonstrating the power of the systematics collections community in many real-world challenges. Work as a community shows the true analytical power of the data that the systematics collections community holds. This power is a key in convincing funding agencies, museum administrators, and decision-makers in general of the importance of systematic collections.

8. CONCLUSIONS

The point of this piece is that computerization is not a prohibitively difficult or expensive endeavor; rather, it is an important step in curating a collection that more than pays for itself in (1) saving time and effort in curatorial activities, (2) improving data quality and removing erroneous elements, and (3) improved funding possibilities and recognition by administrators and decision-makers. Most important is to make some simple decisions, start into the task, and methodically carry it out.

Acknowledgements

This summary is based on long years of work by David A. Vieglais, and several others, in the nascent field of biodiversity informatics. Funding was provided by the U.S. National Science Foundation. Museums included in the Mexican birds summary at the point of the analyses used herein are American Museum of Natural History, Academy of Natural Sciences of Philadelphia, Bell Museum of Natural History, British Museum (Natural History), Carnegie Museum of Natural History, California Academy of Sci-

ences, Canadian Museums of Nature, Denver Museum of Natural History, Delaware Museum of Natural History, Fort Hays State University, Field Museum of Natural History, Iowa State University, University of Kansas Natural History Museum, Los Angeles County Museum of Natural History, Naturhistorisches Museum (Vienna), Louisiana State University Museum of Natural Science, Museum of Comparative Zoology (Harvard University), Moore Laboratory of Zoology (Occidental College), Museum Nationale D'histoire Naturelle (Paris), Museum of Vertebrate Zoology (Berkeley), Museo de Zoología (Facultad de Ciencias, UNAM), University of Nebraska, Royal Ontario Museum, San Diego Natural History Museum, Texas Cooperative Wildlife Collections, University of Arizona, University of British Columbia Museum of Zoology, University of California at Los Angeles, Universidad Michoacana San Nicolás de Hidalgo, United States National Museum of Natural History, Western Foundation of Vertebrate Zoology, and Yale Peabody Museum.

Literature

- BINFORD, L. C. (1989): A distributional survey of the birds of the Mexican state of Oaxaca. *Ornithol. Monogr.* **43**: 1-405.
- NAVARRO-SIGÜENZA, A. G. & PETERSON, A. T. (1999): A restricted type locality for *Cynanthus doubledayi*. *Bull. Brit. Ornithol. Club* **119**: 109-112.
- NAVARRO-SIGÜENZA, A. G. & PETERSON, A. T. (2000): Western Mexico: A significant center of avian endemism and challenge for conservation action. *Cotinga* **14**: 42-46.
- NAVARRO-SIGÜENZA, A. G., PETERSON, A. T. & ESCALANTE-PLIEGO, P. (1992a): New distributional information on Mexican birds. I. The Sierra de Atoyac, Guerrero. *Bull. Brit. Ornithol. Club* **112**: 6-11.
- NAVARRO-SIGÜENZA, A. G., PETERSON, A. T., ESCALANTE-PLIEGO, P. & BENITEZ-DIAZ, H. (1992b): *Cypseloides storeri*: A new species of swift from Mexico. *Wilson Bull.* **104**: 55-64.
- PETERSON, A. T. (1993): Species status of *Geotrygon carrikeri*. *Bull. Brit. Ornithol. Club* **113**: 166-168.
- PETERSON, A. T. (1998): The distribution and type locality of the extinct Slender-billed Grackle, *Quiscalus palustris*. *Bull. Brit. Ornithol. Club* **118**: 119-121.
- PETERSON, A. T. (2001): Predicting species' geographic distributions based on ecological niche modeling. *Condor* **103**: 599-605.
- PETERSON, A. T., FLORES V., O. A., LEON P., L. S., LLORENTE B., J. E., LUIS M., M. A., NAVARRO-SIGÜENZA, A. G., TORRES CH., M. G. & VARGAS F., I. (1993): Conservation priorities in northern Middle America: Moving up in the world. *Biodiversity Letters* **1**: 33-38.
- PETERSON, A. T. & NIETO-MONTES DE OCA, A. (1996): Sympatry in *Abronia* (Squamata: Anguidae) and the problem of Mario del Toro Aviles' specimens. *J. Herpetol.* **30**: 260-262.
- PETERSON, A. T., ESCALONA-SEGURA, G. & GRIFFITH, J. A. (1998a): The birds of northern Central America: A preliminary distributional analysis. *Wilson Bull.* **110**: 534-543.
- PETERSON, A. T., NAVARRO-SIGÜENZA, A. G. & BENITEZ-DIAZ, H. (1998b): The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* **140**: 288-294.
- PETERSON, A. T., EGBERT, S. L., SANCHEZ-CORDERO, V. & PRICE, K. P. (2000): Geographic analysis of conservation priorities using distributional modelling and complementarity: Endemic birds and mammals in Veracruz, Mexico. *Biol. Conserv.* **93**: 85-94.
- PETERSON, A. T., SANCHEZ-CORDERO, V., SOBERON, J., BARTLEY, J., BUDDEMEIER, R. H. & NAVARRO-SIGÜENZA, A. G. (2001): Effects of global climate change on geographic distributions of Mexican Cracidae. *Ecological Modelling*.
- PETERSON, A. T., BALL, L. G. & COHOON, K. C. (2002): Predicting distributions of tropical birds. *Ibis* **144** (Electronic edition): E27-E32.
- PETERSON, A. T., STOCKWELL, D. R. B. & KLUZA, D. A. (in press): Distributional prediction based on ecological niche modeling of primary occurrence data. – In: J. M. SCOTT (ed.) *Predicting species occurrences: Issues of scale and accuracy*. Island Press, Washington, D.C.
- STOCKWELL, D. R. B. (1999): Genetic algorithms II. Pages 123-144 in A. H. FIELDING, editor. *Machine learning methods for ecological applications*. Kluwer Academic Publishers, Boston.
- STOCKWELL, D. R. B. & NOBLE, I. R. (1992): Induction of sets of rules from animal distribution data: A robust and informative method of analysis. *Mathematics and Computers in Simulation* **33**: 385-390.
- Stockwell, D. R. B. & Peters, D. P. (1999): The GARP modelling system: Problems and solutions to automated spatial prediction. *Int. J. Geograph. Informat. Syst.* **13**: 143-158.

A. Townsend PETERSON, Natural History Museum, The University of Kansas, Lawrence, Kansas 66045 USA

Adolfo G. NAVARRO-SIGÜENZA, Museo de Zoología, Facultad de Ciencias, Apartado Postal 70-399, México, D.F. 04510 México

- AND O. D. VEPRINTSEVA. 1996. Natural sound archives: Guidance for recordists and request for cooperation. Pages 474–486 in *Ecology and Evolution of Acoustic Communication in Birds* (D. E. Kroodsma and E. H. Miller, Eds.). Cornell University Press, Ithaca, New York.
- NELSON, D. A., AND S. L. L. GAUNT. 1997. The Borror Laboratory of Bioacoustics (BLB) and the Bioacoustics Research Group at The Ohio State University. *Bioacoustics* 8: 281–286.
- NELSON, D. A., S. L. L. GAUNT, C. L. BRONSON, AND T. J. KLOTH, JR. 2001. Database design for an archive of animal sounds. *IEEE Engineering in Medicine and Biology* 20:76–80.

The Auk 122(3):987–990, 2005
 © The American Ornithologists' Union, 2005.
 Printed in USA.

Ornithology is in a unique position in systematics. Birds are the only major taxon for which more than 99% of species taxa at every point on the surface of the Earth are likely to be known to science (Mayr and Vuilleumier 1983, Peterson 1998). Scientific collections of birds document the distribution and diversity of more than 10,000 species worldwide. Although even these collections are in need of augmentation and improvement (Remsen 1995, Winker 1996, Peterson et al. 1998), data associated with existing specimens constitute a rich source of information about avian distribution and diversity. This resource could serve as the basis for many exciting analyses and insights into the natural history, ecology, systematics, and conservation of birds (Remsen 1995), and as a guide and motivation for further improvement of the specimen basis and information resources.

The need for more efficient access to ornithological data, however, is great. Systematic efforts to document and study avian diversity rely on the specimen record as a critical guide. Biodiversity conservation efforts depend heavily on avian information, as bird distributions can inform conservation planning and prioritization much more completely than other, less well-known taxa. Numerous other applications in natural history, biogeography, ecology, natural

resources management, and even public health also draw insights from avian data (Rappole et al. 2000). This situation thus calls for an efficient system serving accurate ornithological information broadly, both to meet such varied needs and to demonstrate the critical importance of the resource that underlies them.

Presently, such a system does not exist. For example, recent efforts to assemble a list of all specimens of Red Junglefowl (*Gallus gallus*) in natural-history museums in North America and Europe took six and a half months of letter-writing and e-mailing to result in a list of 752 specimens (Peterson and Brisbin 1998). Similarly, efforts to assemble large-scale data sets on migratory bird breeding and wintering areas, necessary for modeling the future distribution of West Nile Virus in North America, were stymied by inefficient access to information and took many months of effort and unnecessary tricks of data manipulation (Peterson et al. 2003).

The technology for such a biodiversity information system nonetheless exists; it was, in fact, developed on the basis of avian data sets, with funding from the National Science Foundation. Subsequently, several efforts have begun assembling such systems across many taxa (see Appendix). Most exciting is that developers of these systems have collaborated to develop a next-generation technology that will meld all these regional efforts into a single, global biodiversity information system—the technology, termed “DiGIR” (distributed generic information retrieval), has won broad acceptance and has been incorporated into many efforts.

Ornithology, with its large quantities of high-quality information regarding an important indicator taxon, has the opportunity to lead this new world of biodiversity informatics. Several other taxonomic communities have already advanced in integrating their data resources via the internet (for examples, see Appendix), and several institutions have already ventured their ornithological data resources in a prototype internet-based distributed system (The Species Analyst, now superseded by ORNIS). Nevertheless, many computerized ornithological data sets remain either unavailable over the internet or available, but not integrated with data sets from other institutions.

Free and open access and data value.—Biodiversity information has traditionally been concentrated in Europe and North America,

even though biodiversity is focused in tropical and subtropical regions. This contrast results from the complexities of the history of scientific exploration, economics, and educational and scientific opportunities. Like biodiversity itself, access to information about biodiversity is unbalanced.

Modern internet technologies make feasible a system in which information resources can be accessed by anyone, anywhere on Earth. The internet provides a medium of information "ow that is limited only by internet access, a barrier that is rapidly disappearing over much of the planet. Hence, the regional imbalances that characterize the current situation can be largely alleviated.

The key point of most debates on the subject of free and open access has been the value of specimen data (Graves 2000). Museum curators know that the information associated with the specimens they curate is valuable, and for that reason they have often guarded such information carefully—the limited budgets at most collections, many of which are in serious financial situations (e.g. recent problems at the Academy of Natural Sciences, Philadelphia), demand that any resource be used wisely. Moreover, resources dedicated to computerization and broad data provision may occur at the expense of specimen care and building the collection itself. However, "valuable" data that are not used yield nothing to the owners or curators of those data.

By contrast, data that are used increase markedly in value. Biodiversity information is too often derived from secondary sources (range maps, field guides, etc.), which both reduces data quality and denies credit to those institutions that house the primary data (often natural-history museums). A system with free and open access to data, however, permits users to access the primary, vouchered information as close to its source as possible. Similar to the marketing strategies of Netscape and Adobe Acrobat, in which providing free and open access is instrumental in building a market share and making a product, such access is key to establishing natural-history museum collections as the premier source of information about biodiversity.

In this sense, the value of data does not decline, but rather increases, as a result of free and open access. That is, as primary ornithological data from specimens become the primary source of information on the distribution of birds, those

data gain value. Furthermore, open access to specimen data results in feedback that leads to higher quality, again increasing the value. By contrast, data for which access is restricted do not benefit to the same extent from analysis, scrutiny, feedback, and interest.

Distributed, not centralized.—A key feature of the information systems under discussion is their distributed nature. Distributed databases may be scattered across regions and countries, but are integrated via the internet. This structure offers distinct advantages: (1) data remain at the owner institution and are usually not centralized; (2) data served can be updated as often as desired, keeping information up-to-the-minute; (3) data ownership is never in question; (4) owner institutions can restrict or limit access as desired (e.g. to limit precision of data regarding distributions of endangered species, to protect rights of investigators regarding publication of works in progress, etc.); and (5) the collaborative nature of the effort is emphasized. Hence, although it required several years of dedicated activity to develop and distribute, this "architecture" makes the idea of providing free and open access to information much more palatable in a number of ways.

Value added.—Serving ornithological information is not a one-way interaction, not just a service to the broader community. Rather, uniting data resources into a single pool allows for several ways of adding value to the primary data. First and foremost, georeferencing locality information becomes much more feasible—because of the redundant nature of localities (specimens from single localities scattered across multiple collections, efficiency of georeferencing work on more densely collected landscapes), such an effort on a collection-by-collection basis is very inefficient. The success of efforts for georeferencing mammal specimen data (Stein and Wieczorek 2004, Wieczorek et al. 2004) is an excellent example. Several additional possibilities—use of ecological niche modeling to detect identification errors, standardization of taxonomic information, and use of collector itineraries to detect date–locality errors—are being developed. All these improvements to data can be repatriated to the owner institutions to improve the base quality of their data sets and information content of the specimens.

Funding potential of community efforts.—A particular advantage of community collaborations is

their excellent potential to leverage funding. The appeal of funding an effort in which all institutions in a community participate is much greater than that of funding an initiative that is based at a single institution. Clear evidence of this potential is the success that several taxonomic groups have had in getting funding for community efforts to integrate data: ichthyology, funded by the National Science Foundation and the Office of Naval Research; mammalogy, funded by the National Science Foundation; and herpetology, funded by the National Science Foundation—summing to more than \$4.5 million in new funding for informatics efforts in scientific collections. These resources would likely not exist without their community basis.

ORNIS and the future.—A fully integrated ornithological information infrastructure has enormous potential, and has now been funded by the National Science Foundation. Approximately $4\text{--}5 \times 10^6$ bird specimens are held in North American museums, and ~80% of those specimens have been committed to participation in ORNIS. Perhaps yet another 4×10^6 bird specimens are held in European museums, and an unknown quantity are held in museums elsewhere in the world ($2\text{--}3 \times 10^6$ more?). Hence, a rough estimate is that on the order of $10\text{--}12 \times 10^6$ bird specimens exist worldwide. If this resource were fully computerized and integrated into a distributed “world museum” of ornithology, the resource would be enormously useful in a broad diversity of applications. Integrating specimen-based data with observational data is enriching the specimen-based information still more: a recent addition to the ORNIS network included 15×10^6 observational records from several projects based at the Cornell Laboratory of Ornithology.

At present, much information about birds is drawn from secondary sources. Conservation organizations prepare secondary information resources (lists of endangered species, distributional summaries, etc.). Field guides synthesize information into range summaries and distribution maps. Other resources are assembled solely on the basis of observational information, which lacks vouchering and can be unreliable in some circumstances (Phillips 1986). These secondary resources are too often used as the basis for answering important questions about birds.

Why are specimen data—the ultimate “library of life” information resource for biodiversity—

not already the primary information resource for birds? The answer lies in the difficult and inefficient access that has characterized this resource. Simply, the data are not used because they are hard to access. As ornithology provides better and more efficient access to specimen data resources—via ORNIS and related solutions, and their descendents—the user base will grow. Only in this way can avian collections get the key recognition and support they deserve and need.—A. TOWNSEND PETERSON, *Natural History Museum and Biodiversity Research Center, University of Kansas, Lawrence, Kansas 66045, USA (e-mail: town@ku.edu)* and CARLA CICERO AND JOHN WIECZOREK, *Museum of Vertebrate Zoology, University of California, Berkeley, California 94720, USA.*

LITERATURE CITED

- GRAVES, G. R. 2000. Costs and benefits of web access to museum data. *Trends in Ecology and Evolution* 15:374.
- MAOR, E., AND F. VUILLEUMIER. 1983. New species of birds described from 1966 to 1975. *Journal für Ornithologie* 124:217–232.
- PETERSON, A. T. 1998. New species and new species limits in birds. *Auk* 115:555–558.
- PETERSON, A. T., AND I. L. BRISSON. 1998. Genetic endangerment of wild Red Junglefowl *Gallus gallus*? *Bird Conservation International* 8: 387–394.
- PETERSON, A. T., A. G. NAVARRO-SIENZA, AND H. BENOITEZ-DAZ. 1998. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* 140:288–294.
- PETERSON, A. T., D. A. VIEILAIS, AND J. ANDREASEN. 2003. Migratory birds as critical transport vectors for West Nile Virus in North America. *Vector Borne and Zoonotic Diseases* 3:39–50.
- PHILLIPS, A. R. 1986. *The Known Birds of North and Middle America. Part I.* Privately published, Denver, Colorado.
- RAPPOLE, J., S. R. DERRICKSON, AND Z. HUOLEK. 2000. Migratory birds and spread of West Nile virus in the Western Hemisphere. *Emerging Infectious Diseases* 6:319–328.
- REMSEN, J. V., JR. 1995. The importance of continued collecting of bird specimens to ornithology and bird conservation. *Bird Conservation International* 5:145–180.

- STEIN, B., AND J. WIECZOREK. 2004. Mammals of the World: MANIS as an example of data integration in a distributed network environment. *Biodiversity Informatics*, no. 4. [Online.] Available at jbi.nhm.ku.edu/viewarticle.php?id=11.
- WIECZOREK, J., Q. GUO, AND R. J. HIJMANS. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18: 745–767.
- WINKER, K. 1996. The crumbling infrastructure of biodiversity: The avian example. *Conservation Biology* 10:703–707.

APPENDIX

The following are websites for efforts to assemble biodiversity information systems: MaNIS (elib.cs.berkeley.edu/manis/); HerpNet (www.herpNet.org/); Global Biodiversity Information Facility (www.gbif.net/); Red Mundial para la Información de la Biodiversidad (www.conabio.gob.mx/remib/doctos/remib_esp.htm); Virtual Australian Herbarium (www.rbgsyd.gov.au/HISCOM/Virtualherb/virtualherbarium.html); SpeciesLink (www.cria.org.br/projetos/); European Natural History Specimen Information Network (www.nhm.ac.uk/science/rco/enhsin/). For information on distributed generic information retrieval (DiGIR), go to digir.sourceforge.net/.

For examples of taxonomic data resources on the internet, see www.speciesanalyst.net/fishnet/ (ichthyology); elib.cs.berkeley.edu/manis/ (mammalogy); www.herpNet.org (herpetology). On efforts for georeferencing mammal specimen data, see elib.cs.berkeley.edu/manis/. The ORNIS website is at ornisnet.org.

both traditional and nontraditional uses of these shared research resources. For example, advances in analytical chemistry have enabled researchers to obtain data on heavy-metal contaminants and diets from a single feather. Future technological advances will increase nontraditional use of specimens, and two areas of rapid growth at present are in contaminant and stable-isotope studies. We address these developments and their implications for bird collections.

Contaminants.—Retrospective contaminant studies of the 1960s and 1970s premiered a new and important use of specimens. One of the first studies to use bird specimens in contaminant research documented a 10- to 20-fold increase in feather mercury among seed-eaters and raptors after the introduction of alkyl-mercury seed dressings (fungicides) in Europe in the 1940s (Berg et al. 1966). That research led to the banning of those seed treatments, and subsequent retrospective analyses using specimens confirmed the effect of alkyl-mercury fungicides by documenting the decline of mercury concentrations in feathers after the ban (Westermarck et al. 1975). Probably the best-known use of museum specimens in retrospective research documented eggshell thinning in raptors following the introduction of DDT in 1947 (Ratcliffe 1967, Hickey and Anderson 1968). These studies and others (see Kiff 2005) contributed to the eventual ban of DDT in many countries.

Researchers have documented high levels of contaminants in the biota of undeveloped regions, citing the global distribution of pollutants as the cause (Arctic Monitoring and Assessment Programme 1998). As global contaminant burdens increase, spatially and temporally distributed biological samples are needed to document changing contaminant levels. Archived avian specimens can document levels of heavy metals, because heavy metals bind to feather keratin at the time of growth (Crewther et al. 1965). Archived specimens were used to document increases in mercury pollution in several avian food webs (Appelquist et al. 1985; Thompson et al. 1992, 1993). Time series of archived seabirds were also used to document increases in feather mercury concentrations in two avian food webs over the past 100 years, which were correlated with anthropogenic inputs (Monteiro and Furness 1997, Thompson et al. 1998).

Use of Bird Collections in Contaminant and Stable-isotope Studies.—Preserved biological specimens are increasingly providing source material for research that is moving beyond traditional questions in collections-based studies. Technological advances are facilitating

RECOMMENDATION ON OPEN ACCESS TO BIODIVERSITY DATA (Adopted by GBIF Governing Board on 16.01.06)

The Global Biodiversity Information Facility (GBIF) Governing Board -- representing 47 countries, 31 international organizations and the Secretariat on the Convention of Biological Diversity - hereby recommends that research councils, other funding agencies and private foundations:

- **Promote that proposals for funding for biodiversity research include a plan for the maintenance and sharing of the digital biodiversity data generated in proposed projects;**
- **Promote that species and specimen level data and associated metadata that are generated in funded projects are made publicly available through mechanisms cooperating with GBIF, within a specified period after completion of the supported research.**

Rationale:

Many research projects generate biodiversity data sets that are relevant for the wider scientific community, government natural resource managers, policy makers, and the public. Because data sharing now requires small marginal costs compared to the full research costs that generate the data, it is wise to allow for further shared use of these data to benefit the widest possible range of users.

The UN Convention on Biological Diversity has called for more data and information for the effective implementation of its workplans, and the key goals of conservation, sustainable use and the sharing of benefits from the utilization of genetic resources. The World Summit on the Information Society (WSIS) in Geneva in December 2003 strongly affirmed the principle of “universal access with equal opportunities for all to scientific knowledge and the creation and dissemination of scientific and technical information.”

Two of the goals of GBIF are to bring together data for multiple uses, and to find incentives and mechanisms to make data freely available as quickly and effectively as possible. These goals underlie the recommendations made here. GBIF’s initial focus is to make available as much data on species and specimens as possible, and to this end it has developed standards and tools. In the coming years, other biodiversity data elements will be built into the GBIF infrastructure. Indeed, GBIF’s ability to build tools and bring together information led the CBD Conference of the Parties to recognize the potential lead role of the GBIF in facilitating its work in the Global Taxonomy Initiative (COP 6) and Inlands Waters (COP 7).

The advantages of free and open data sharing have been documented (Arzberger et al. 2004) and brought together in the collaborative Conservation Commons (www.conservationcommons.org):

- Sharing data is good scientific practice and is necessary for the advancement of science, public awareness and education;
- Expanded access to data sources could impressively increase the value to taxpayers of the more than \$650 billion spent annually by governments on all research disciplines (Science, Technology and Innovation for the 21st

Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique);

- The openness of science stimulates and facilitates creativity;
- Open access to data enables greater accountability to funding sources as quality, reliability, productivity and use of data are enhanced with public utilization and review.

Requirements for open access to data (e.g. National Institutes of Health, 2003; National Science Foundation 2001) signal the importance of data sharing to science and to decision-making, as well as to the long-term benefits to society and the environment, while respecting the right of scientists to publish on their data before releasing it for use by others.

REFERENCES

Arzberger P., P. Schroeder, A. Beaulieu, G. Bowker, K. Casey, L. Laaksonen, D. Moorman, P. Uhler, P. Wouters (2004). Promoting Access to Public Research Data for Scientific, Economic, and Social Development. *Data Science Journal* 3, 135-152. Retrieved 2005.06.07 from http://journals.eecs.qub.ac.uk/codata/Journal/contents/3_04/3_04pdfs/DS377.pdf

ENBI – GBIF Digitisation Workshop Participants' Statement on Free and Open Data Access (January 2005). Retrieved 2005.06.07 from http://circa.gbif.net/Public/irc/gbif/pr/library?l=/access_statement/EN_1.0_&a=i.

Froese, R and R. Reyes Jr. (2003). Use them or lose them: the need to make collection databases publicly available. In A. Legakis, S. Sfenthourakis, R. Polymeri and M. Thessalou-Legaki (eds.) *Proceedings of the 18th International Congress of Zoology*, 585-591. Retrieved 2006 01 16 from <http://filaman.ifm-geomar.de/ifm-geomar/rfroese/UseOrLose.doc>

Froese, R., D. Lloris and S. Opitz. (2004). The need to make scientific data publicly available; concerns and possible solutions. In M.L.D. Palomares, B. Samb, T. Diouf, J.M. Vakily and D. Pauly (eds.) *Fish Biodiversity: Local studies as basis for global inferences. ACP-EU Fisheries Research Report* 14, 268-271. Retrieved 2006 01 16 from <http://filaman.uni-kiel.de/ifm-geomar/rfroese/ConcernsDataowners.pdf>

Moritz, Thomas Daniel (2004) Conservation Partnerships in the Commons? *Museum* 56:4, 24-31. Retrieved 2005.10.12 from http://www.eco-index.org/search/pdfs/moritz_english.pdf

National Institutes of Health. 2003. Final NIH statement on sharing research data. Retrieved 2005.06.07 from <http://grants2.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>

National Science Foundation Grant General Conditions. 2001. Article 36. Sharing of Findings, Data, and Other Research Products. Page 17. Retrieved 2005.06.07 from <http://www.nsf.gov/pubs/2001/gc101/gc101rev1.pdf>.

Ocean Biodiversity Informatics conference statement, Hamburg, 1 December 2004. Retrieved 2005.06.07 from <http://www.vliz.be/obi/statement.php>.

OECD Follow Up Group on Issues of Access to Publicly Funded Research Data. *Promoting Access to Public Research Data for Scientific, Economic, and Social Development: Final Report* March 2003.

Science, Technology and Innovation for the 21st Century. Meeting of the OECD Committee for Scientific and Technological Policy at Ministerial Level, 29-30 January 2004 - Final Communique. Retrieved 2005.06.07 from http://www.oecd.org/document/15/0,2340,en_2649_201185_25998799_1_1_1_1,00.html.

Website of the UN Convention on Biological Diversity, www.biodiv.org.

**Report of the *Pro Bono* Legal Expert Group Meeting
Global Biodiversity Information Facility
18 – 19 September 2006**



Legal Expert Group (ProLEG) to identify legal issues of importance to GBIF and to analyze them and provide recommendations on how to address them. The Group consisted of legal and other experts from Africa, Asia, Europe, Latin America and North America, who provided advice in their personal expert capacity and not as representative of their respective institutions of employment.

The ProLEG met once outside Copenhagen on 18-19 September 2006 to consider the issues raised by GBIF and to write the first draft of this report. The list of ProLEG members and GBIF Secretariat participants in this meeting is provided in Appendix A. The report was completed subsequently by e-mail consultations. It identifies the key issues that the Group believes GBIF needs to address, provides conclusions in each issue area, and establishes recommendations in response to the conclusions.¹ Unless specifically stated otherwise, all references to GBIF in this report refers to the GBIF Secretariat and to all organizations involved in GBIF, including its Participant Members, Affiliated Members, and data providers.

ISSUES, CONCLUSIONS, AND RECOMMENDATIONS

1. The public, cooperative nature of GBIF and its users.

GBIF and the users of its portal (www.gbif.net) are engaged predominantly in public scientific, educational, and not-for-profit activities. The GBIF portal constitutes a central interface between GBIF's data providers and users, which therefore raises certain expectations and requirements on both sides that require attention by all parties involved in the project.

As stated in the GBIF Memorandum of Understanding (MoU) and the data sharing agreement, the presumption is that data providers will, to the greatest extent possible, make their data freely and openly available through the GBIF portal, subject to a user requirement of due attribution for the data source. Moreover, the primary biodiversity data that are the focal point of GBIF services have broad public-interest value.

Although the organizations and activities associated with GBIF are largely public and not-for-profit, there are a number of potentially significant legal consequences that may arise and should be considered expressly by GBIF management. GBIF, like all other

¹ This report, however, should not be construed as providing legal advice. GBIF and its participating organizations should consult with their legal counsel regarding the implementation of any recommendations or related activities that may involve potential legal liabilities.

individuals and organizations, is subject to the well-established legal principle that ignorance of the law is no excuse.

Recommendation 1: *Considering that the mandate and purpose of GBIF is to promote the sharing of primary biodiversity data freely and openly, GBIF should seek to rely upon and use, as much as possible, the practices norms, and policies of public science to guide its activities and avoid using legalistic solutions and enforcement mechanisms.*

2. Legal status of different information products made available through the GBIF portal.

As is the case with all data and databases and other types of information products, such as literature and software, their legal status is subject to a range of public laws (treaties, legislation, regulations and their interpretation in different jurisdictions by the judiciary) and increasingly to private law (licensing agreements and contracts) in the online environment. Although a detailed discussion of these legal sources is beyond the scope of this report, a basic outline of these sources and their relevance to the different information made available through GBIF is useful for framing the discussion that follows. A much more comprehensive treatment of the relevant legal sources, concepts, and application to GBIF data and other information products is provided in the report commissioned by GBIF in 2004, “An analysis of the implications of intellectual property rights (IPR) on the *Global Biodiversity Information Facility (GBIF)*”, by Manuel Ruiz Muller, which may be found on the GBIF Web site at: <http://www.gbif.org/prog/ocb/iprmtg/IPRanalysis.pdf>

The most important public law sources are intellectual property (IP) statutes such as copyright and, in those jurisdictions that have it, database protection legislation. Patent law may be relevant to certain data, such as genomic or proteomic, but such data constitute only a small percentage of data that may be made available through GBIF and the legal aspects of that information are the bailiwick of the source organizations. In this regard, it should be noted that patents do not protect data per se, but only the practice, making, or selling of inventions. Thus, data content or transfers cannot, by themselves, infringe a patent, although specific uses or applications of those data may.

Other legislation and regulations applicable to biodiversity research activities include laws governing biodiversity samples and related information. For example, some developing countries are enacting new laws that may limit disclosure of data or information about biological resources originating from these countries. Consistent with the Convention on Biodiversity, over 30 countries have enacted or introduced legislation that regulates access to the genetic resources within their jurisdictions, and related benefit sharing. In some cases, such legislation also regulates the access to and use of data and information about those resources for commercial purposes (e.g., *The Biological Diversity Act, 2002* [No. 18 of 2003], Ministry of Law and Justice, India, available at: http://www.envfor.nic.in/divisions/biodiv/act/bio_div_act_2002.pdf). Although these

laws are not directly applicable to GBIF's data activities at present, it is not clear how they may affect GBIF in the future.

Licenses and contracts are increasingly used for the dissemination of digital information from the owner to the user. The validity of such private law instruments depends on whether they meet all the legal criteria in the jurisdiction(s) in question, as discussed further below.

With regard to the information products that are made available through the GBIF portal, an important legal distinction needs to be made between facts and compilation of those facts into substantial data sets or entire databases. Under both copyright and *sui generis* database protection statutes, individual facts are in the public domain. Under copyright law, moreover, all the factual data remain in the public domain and only the original and creative selection and arrangement of those data may be subject to copyright protection. However, under the database protection law in the EU and some other countries, "substantial parts" of a database are protected if they resulted from substantial investment by the rights holder. Thus, in these jurisdictions many databases are presumably subject to protection under database protection laws and discrete subsets of data also may be.

These differences in IP law in different jurisdictions may affect the applicability of license agreements and of restrictions on data use, including attribution and non-commercial restrictions imposed by data providers. Licenses are applicable when relevant law protects the data or databases (for example, under copyright or *sui generis* database rights). In contrast, contracts can impose conditions on use, even without statutory rights. However, the use of contracts for this purpose is limited in practice by contract formation principles and validity issues, as discussed further in Section 4, below.

<p><i>Recommendation 2: Consistent with Recommendation 1 and the relevant statutory law, GBIF should impose the least possible restrictions and obligations on users.</i></p>

3. Seeking permission from originating data sources by data providers.

Most data collections of providers that are made available via GBIF are compiled from multiple sources, and some portions of these collections may be of unknown origin or legal status. This situation can undermine the ability of data providers to warrant that they have made all the necessary agreements with the original owners of the data in order to make the data available online through the GBIF portal. Nevertheless, data providers should not neglect their responsibility to seek permission for providing data. At a minimum, the data providers should warrant that they have made reasonable attempts to locate and seek the consent of the original sources of the specimens or of the data sources, and to provide due attribution to them. Failure to obtain the requisite permissions and to give due attributions can result in damaging negative publicity to the data provider and GBIF, and undermine their reputation and ability to pursue their objectives. Moreover, the issue here is not only a question of legal conformity, but the quality and

reliability of the data is in doubt when you cannot specify their origin and have access to the source.

Recommendation 3: GBIF should consider revising Clause 1.3 of the GBIF Data Sharing Agreement as follows: “The data provider has made reasonable efforts to ensure that the original owner(s) of the data have agreed that the data may be made available on the GBIF Web site. The data provider also should disclose existing information about the origin of the data in order to allow appropriate recognition and attribution of the original source.”

4. Attribution requirement on users.

As alluded to in the previous section, appropriate attribution is an important benefit to the original data sources and to the providers of biodiversity data. It provides recognition of work and reputational benefits to the organizations (and individuals) participating in GBIF. It also contributes to the transparency of the activity and supports good scientific norms and research processes. Developing countries are particularly concerned that due attribution is not being given to data accessed from them. Attribution thus not only constitutes an equitable condition of free and open data dissemination and reuse, but provides one of the few incentives for the data sources and providers to continue to make their data freely and openly available.

Based on the legal status of data and databases under IP law in different jurisdictions, however, there are substantial problems with the legal validity and enforcement of the attribution requirement for data. The GBIF contractual requirement regarding due attribution by users is defective for individuals or entities in jurisdictions that do not grant statutory database protection such as the E.U. Database Directive for the following reasons, when there is:

- No underlying statutory enforcement of such a right;
- Uncertain validity of “click-through” licenses online; or
- Insufficient consideration given by each party (i.e., no *quid pro quo*).

Acknowledgement may be enforceable for substantial data sets and entire databases in jurisdictions that have database protection legislation, since such laws confer exclusive property rights in substantial parts (measured quantitatively or qualitatively) of data collections resulting from substantial investment. Acknowledgement may be enforceable under copyright law as well, if the database is copyrightable.

However, even under these conditions, the license would be valid only for the signing party and not for other third parties (who lack privity of contract). Moreover, as noted in Section 2 above, there is no legal basis for requiring or enforcing the attribution of individual facts (or insubstantial amounts of data) by re-users of the data, which most likely would eliminate most users of GBIF data from the ambit of the attribution requirement, even in those jurisdictions that have enacted database protection legislation.

The attribution “requirement” thus should be viewed and treated by GBIF as a request to follow good scientific practice and generally accepted normative conduct.

Recommendation 4: GBIF should continue to include attribution as a condition of the use of the data through its portal in order to encourage such normative behavior by the data users.

5. *Non-commercial requirements on users.*

For the same reasons discussed above, non-commercial limitations on the use of data created by licensing agreements in the absence of underlying statutory protection could be defective. Moreover, different providers may have different conceptions of “non-commercial use” or “commercial use” (or even define the terms differently in their contracts). For example, “commercial use” may be defined as being dependent upon an entity’s status (i.e., non-profit vs. for-profit) or on the type of activity (i.e., whether payment is received for the reuse of the data). Data providers may have further variations on the scope of their own definitions of what constitutes commercial and non-commercial uses. Such varying definitions would be difficult to implement or communicate in a standard GBIF agreement. Existing Creative Commons licenses (using database protection rights for their enforcement) therefore may be inappropriate because of potential inconsistencies between the Creative Commons definition of “non-commercial” and the requirements of GBIF (or its data providers). In any case, restrictions or obligations should not be imposed by contract on individual facts which are not protected by any relevant law or norm.

Recommendation 5a: GBIF should continue to work with its data providers to promote its free and open data access policy, subject only to appropriate attribution.

Recommendation 5b: For those data providers that require restrictions on commercial reuse of their data, the development of a standardized licensing mechanism similar to the Creative Commons licenses could be appropriate. This is an area that requires further investigation by GBIF and legal experts on data and information licensing in different jurisdictions.

6. *Enforcement of terms and conditions on users.*

Legal enforcement of terms and conditions on the use of data would involve negative adversarial aspects and substantial costs associated with monitoring the uses and asserting the rights. Enforcement mechanisms can include various approaches, including legal (e.g., threatening legal action through cease and desist letters, or filing law suits), technological (e.g., use of Trusted Platform Management - TPM, Digital Rights Management - DRM, or persistent, unique identifiers), and normative (e.g., use of scientific ethics and publicizing of transgressions through the “mobilization of shame”). For reasons discussed above, legal approaches are not appropriate for enforcing GBIF’s

attribution requirement and may be appropriate for challenging commercial abuses only in exceptional circumstances involving wholesale infringement and misappropriation of large collections.

TPM and DRM tools should not be used if they undermine the primary objective of free and open access online, or impose undue costs or new obligations on providers and users. Persistent and unique identifiers are useful and appropriate and are being investigated by GBIF at this time. These tools can help track various uses and users automatically. They may have substantial costs associated with their adoption and raise some privacy concerns that can be contrary to the values promoted by GBIF, however. Users also may be blocked from access. Such problems may be especially burdensome to data providers and users in developing countries. In any case, trust and transparency must remain the main values underlying GBIF policy.

Recommendation 6a: GBIF should consider normative enforcement methods that rely on the promotion of ethical scientific practice, good will, and peer pressure as a soft and low-cost alternative, and in conformity with the values and objectives promoted by the organization.

Recommendation 6b: Publicizing of inappropriate behavior related to persistent non-compliance with important terms and conditions of data use may be considered, but only in consultation with legal counsel.

7. Barriers to the addition of new GBIF Participants or of disclosing data through the GBIF portal.

The characterization of barriers, including legal barriers, is incomplete and their effect on GBIF participation or on data availability is not fully known or understood. Trade-related benefits, protection of traditional indigenous knowledge, biological conservation, management of national natural resources, laws and policies based on national security, and other considerations based on perceived national interests may come into play. Some of these constraints, whether reasonable or not, may apply even to the primary biodiversity data that constitute the basic content on the GBIF portal. Economic and technological asymmetries among nations can exacerbate the perceptions of these problems.

There was a diversity of viewpoints among the ProLEG members about more specific requirements or terms that GBIF might incorporate into its data use and data sharing agreements, including: disclosure requirements--disclosure of the origin of the source of the information; prior informed consent--permission of the original owner to provide the information; benefit sharing from use of data from developing countries; technology transfer and technical assistance, particularly for overcoming technological barriers to accessing and using data; and privacy concerns. There was no consensus about these

specific approaches, however, so they are not offered as formal recommendations of the committee.

Recommendation 7: GBIF should continue to develop a strategy for dealing with the barriers perceived by potential participants, consistent with its fundamental data access and use principles.

8. Risk Management.

There are a number of potential, though remote, liabilities for GBIF (and its other member organizations) arising from public disclosure of erroneous and harmful data, illegal data disclosure, or negligence.

Recommendation 8a: GBIF should consult with legal counsel to determine relevant liabilities and whether it would be prudent to obtain insurance to cover such risks.

Recommendation 8b: Clause 1.9 of the Data Sharing Agreement should be merged into clause 1.10, because it is largely redundant. The new clause should read: “GBIF Secretariat is not liable or responsible, nor are its employees or contractors, for the data contents or their use; or for any loss, damage, claim, cost or expense however it may arise, from an inability to use the GBIF network.”

APPENDIX A

GBIF *Pro Bono* Legal Expert Group Meeting 18-19 September 2006 Gentofte, Denmark

List of Participants

Paul Uhlir (Chair)
Director, International Scientific and Technical Information Programs
National Research Council
Washington, DC, USA

Paul Asimwe
Sipi Law Associates
Advocates & Legal Consultants
Kampala, Uganda

Daniele Bourcier
Director of Research
CERSA-CNRSA
Berlin, Germany

Philippe Desmeth
Service Public Federal de Programmation Politique Scientifique
Belgian Federal Science Policy Office
Brussels, Belgium

Anitha Ramanna
Lecturer, Department of Politics & Public Administration
University of Pune
Pune, India

Manuel Ruíz
Sociedad Peruana de Derecho Ambiental (SPDA)
Lima, Peru

Thinh Nguyen
Counsel, Science Commons
Cambridge, MA, USA

China Williams
CBD Unit
Royal Botanic Gardens, Kew
Richmond, Surrey, UK

Convention on Biological Diversity

Article 17. Exchange of Information

1. The Contracting Parties shall facilitate the exchange of information, from all publicly available sources, relevant to the conservation and sustainable use of biological diversity, taking into account the special needs of developing countries.
2. Such exchange of information shall include exchange of results of technical, scientific and socio-economic research, as well as information on training and surveying programmes, specialized knowledge, indigenous and traditional knowledge as such and in combination with the technologies referred to in Article 16, paragraph 1. It shall also, where feasible, include repatriation of information.

Recommendation

Approved by the GBIF Governing Board 17 October 2007

GBIF's contribution to CBD's Article 17: Exchange of Information

Background:

Due to the immense potential of 'virtual collections' to enhance scientific knowledge and improve conservation and sustainable use of biodiversity, Parties to international treaties like the Convention on Biological Diversity (CBD) have repeatedly emphasised the importance of repatriation of information to countries of origin (e.g. CBD Art. 17 paragraph 2); and decisions CBD COP III/10 and COP IV/1.D have stressed the role that scientific institutions - particularly in the developed world - can play in mobilising these data.

Recommendation: Based on the Open Access recommendation adopted by the GBIF GB11, the GBIF Governing Board adopts the following two resolutions on mobilising biodiversity data and sharing data with countries of origin:

1. The GBIF Governing Board recommends that natural history institutions housing biodiversity materials from other countries:
 - **Ensure that species and specimen-level data and associated metadata be digitised and made openly and publicly available through mechanisms cooperating with GBIF.**
2. The GBIF Governing Board recommends that research organisations, research councils, governmental, non-governmental organisations, international agencies, funding agencies and private foundations around the world:
 - **Provide funding for research, capacity building, training and other relevant activities that include the digitisation and open dissemination of species and specimen-level data collected beyond their national territories, in accordance with GBIF-mediated standards and protocols.**

GBIF Data Sharing Agreement (most recent update, as of 23 Nov 2007)

Background

The goals and principles of making biodiversity data openly and universally available have been defined in the Memorandum of Understanding on GBIF, paragraph 8 (see the relevant excerpts in the attached Annex).

The Participants who have signed the MoU have expressed their willingness to make biodiversity data available through their nodes to foster scientific research development internationally and to support the public use of these data.

Data providers often participate in several data sharing arrangements at different levels (thematic, community, national, global).

GBIF data sharing should take place within a framework of due attribution.

Therefore, when registering their services with GBIF, the data providers agree as follows:

1. Data Sharing Agreements

1. Biodiversity data accessible via the GBIF network are openly and universally available to all users within the framework of the GBIF Data Use Agreement and with the terms and conditions that the data provider has identified in its metadata.
2. GBIF does not assert any intellectual property rights in the data that is made available through its network.
3. The data provider warrants that they have made the necessary agreements with the original owners of the data that it can make the data available through GBIF network.
4. The data provider makes reasonable efforts to ensure that the data they serve are accurate.
5. Responsibility regarding the restriction of access to sensitive data resides with the data provider.
6. The data provider includes stable and unique identifiers in their data so that the owner of the data is known and for other necessary purposes.
7. GBIF Secretariat may cache a copy and serve full or partial data further to other users together with the terms and conditions for use set by the data provider. Queries of such data through the GBIF Secretariat are reported to the data provider.
8. Data providers are endorsed by a GBIF Participant, if applicable, before their metadata is made available by the GBIF Secretariat.
9. GBIF Secretariat is not liable or responsible, nor are its employees or contractors, for the data contents or their use; or for any loss, damage, claim, cost or expense however it may arise, from an inability to use the GBIF network.

2. Service Levels

GBIF Secretariat

1. Services provided by the GBIF Secretariat are managed in accordance with the GBIF Work Programme.
2. GBIF Secretariat's service provision includes software components and updates, interfaces, indexing and registry services, helpdesk, and training to assist the Participants to maintain Internet portals.

GBIF Participants

3. GBIF Participants keep the GBIF Secretariat informed of their contact and service information.
4. GBIF Participants maintain services that enable new and existing data providers in their domain to be integrated within GBIF network, and the data owners be identified, as appropriate.

3. Definitions

- GBIF Participant: Signatory of the GBIF-establishing Memorandum of Understanding (MoU).
- GBIF Secretariat: Legal entity empowered by the GBIF Participants to enter into contracts, execute the Work Programme, and maintain the central services for the GBIF network.
- GBIF network: The infrastructure consisting of the central services of the GBIF Secretariat, Participant nodes and data providers. Making data available through GBIF network means registering and advertising the pertinent services via the GBIF central services.
- Participant Node: A coordinating agency or institution designated or established by a GBIF Participant that promotes, coordinates, and facilitates biodiversity data sharing activities by working with data providers in its domain and using GBIF's central services.
- Biodiversity Data: Primary data on specimens, observations, names, taxonomic concepts, and sites, and other data on biological diversity.
- Dataset: A compilation of related data records
- Metadata: Data that describes the attributes of datasets and their constituent records.
- Data sharing: The process of and agreements for making data freely and universally available on the Internet.
- Data provider: A custodian of data who makes it accessible via the Internet. This may or may not be the data owner. If not, they will have declared to GBIF that they have permission to make the data available.
- User: Anyone who uses the Internet to access data through the GBIF network.
- Owner of data: The legal entity possessing the rights resulting from the act of creating a digital record. The record may be a product derived from another, possibly non-digital product, which may affect the rights.
- Sensitive data: Any data that the data provider does not want to make available, for example the precise localities for threatened species.

[Back to Introduction](#)

GBIF Data Use Agreement (most recent update, as of 23 Nov 2007)

Background

The goals and principles of making biodiversity data openly and universally available have been defined in the Memorandum of Understanding on GBIF, paragraph 8 (see the relevant excerpts in the attached Annex).

The Participants who have signed the MoU have expressed their willingness to make biodiversity data available through their nodes to foster scientific research development internationally and to support the public use of these data.

GBIF data sharing should take place within a framework of due attribution.

Therefore, using data available through the GBIF network requires agreeing with the following:

1. Data Use Agreements

1. The quality and completeness of data cannot be guaranteed. Users employ these data at their own risk.
2. Users shall respect restrictions of access to sensitive data.
3. In order to make attribution of use for owners of the data possible, the identifier of ownership of data must be retained with every data record.
4. Users must publicly acknowledge, in conjunction with the use of the data, the data providers whose biodiversity data they have used. Data providers may require additional attribution of specific collections within their institution.
5. Users must comply with additional terms and conditions of use set by the data provider. Where these exist they will be available through the metadata associated with the data.

2. Citing Data

Use the following format to cite data retrieved from the GBIF network:

Biodiversity occurrence data provided by: (Accessed through GBIF Data Portal, www.gbif.net, YYYY-MM-DD)

For example:

Biodiversity occurrence data provided by: Field Museum of Natural History, Museum of Vertebrate Zoology, University of Washington Burke Museum, and University of Turku (Accessed through GBIF Data Portal, www.gbif.net, 2007-02-22)

3. Definitions

- GBIF Participant: Signatory of the GBIF-establishing Memorandum of Understanding (MoU).
- GBIF Secretariat: Legal entity empowered by the GBIF Participants to enter into contracts, execute the Work Programme, and maintain the central services for the GBIF network.
- GBIF network: The infrastructure consisting of the central services of the GBIF Secretariat, Participant nodes and data providers. Making data available through GBIF network means registering and advertising the pertinent services via the GBIF central services.
- Participant Node: A coordinating agency or institution designated or established by a GBIF Participant that promotes, coordinates, and facilitates biodiversity data sharing activities by working with data providers in its domain and using GBIF's central services.
- Biodiversity Data: Primary data on specimens, observations, names, taxonomic concepts, and sites, and other data on biological diversity.
- Dataset: A compilation of related data records
- Metadata: Data that describes the attributes of datasets and their constituent records.
- Data sharing: The process of and agreements for making data freely and universally available on the Internet.
- Data provider: A custodian of data who makes it accessible via the Internet. This may or may not be the data owner. If not, they will have declared to GBIF that they have permission to make the data available.
- User: Anyone who uses the Internet to access data through the GBIF network.
- Owner of data: The legal entity possessing the rights resulting from the act of creating a digital record. The record may be a product derived from another, possibly non-digital product, which may affect the rights.
- Sensitive data: Any data that the data provider does not want to make available, for example the precise localities for threatened species.

Also see the [GBIF Data Sharing Agreement](#) for the data providers.