

PRINCIPLES OF DATA QUALITY

Arthur D. Chapman¹

*Although most data gathering disciples treat error as an embarrassing issue to be expunged, the error inherent in [spatial] data deserves closer attention and public understanding ...because error provides a critical component in judging fitness for use.
(Chrisman 1991).*



¹ Australian Biodiversity Information Services
PO Box 7491, Toowoomba South, Qld, Australia
email: papers.digit@gbif.org

© 2005, Global Biodiversity Information Facility

Material in this publication is free to use, with proper attribution. Recommended citation format:

Chapman, A. D. 2005. *Principles of Data Quality*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

This paper was commissioned from Arthur Chapman in 2004 by the GBIF DIGIT programme to highlight the importance of data quality as it relates to primary species occurrence data. Our understanding of these issues and the tools available for facilitating error checking and cleaning is rapidly evolving. As a result we see this paper as an interim discussion of the topics as they stood in 2004. Therefore, we expect there will be future versions of this document and would appreciate the data provider and user communities' input.

Comments and suggestions can be submitted to:

Larry Speers
Senior Programme Officer
Digitization of Natural History Collections
Global Biodiversity Information Facility
Universitetsparken 15
2100 Copenhagen Ø
Denmark
E-mail: lspeers@gbif.org

and

Arthur Chapman
Australian Biodiversity Information Services
PO Box 7491, Toowoomba South
Queensland 4352
Australia
E-mail: papers.digit@gbif.org

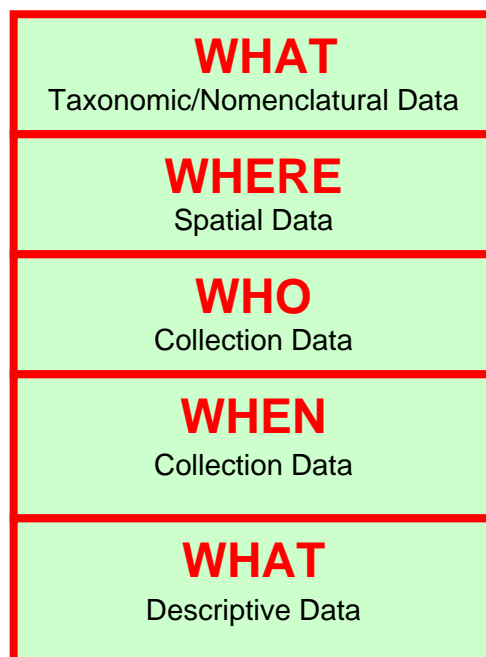
July 2005

Cover image © Per de Place Bjørn 2005
Amata phegea (Linnaeus 1758)

Contents

Introduction	1
Definitions	3
Principles of Data Quality	8
Taxonomic and Nomenclatural Data.....	21
Spatial Data	25
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	45
Conclusion.....	50
Acknowledgements	51
References	52
Index.....	57

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).

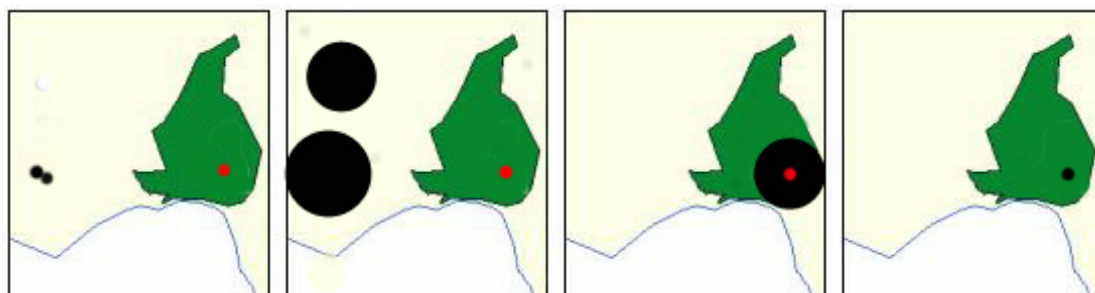


Fig. 1. Shows the differences between accuracy and precision in a spatial context. The red spots shows the true location, the black spots, represent the locations as reported by a collector.

- 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.

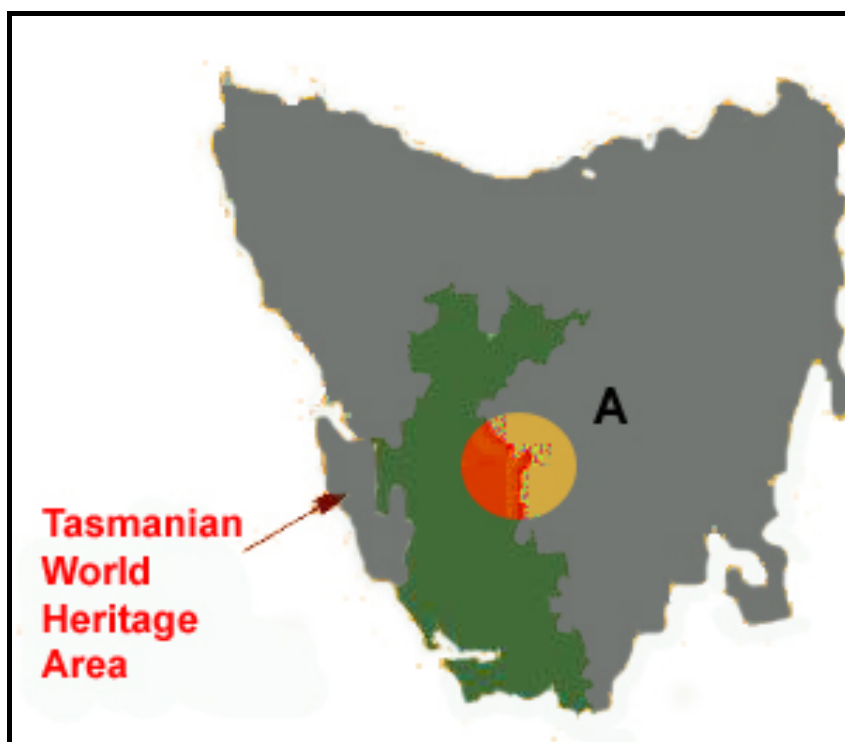


Fig. 2. Map of Tasmania, Australia, showing a record (A) collected with an accuracy of 0.5° (ca. 50 km) as shown by circle. The footprint area of possible collection (determined using the accuracy value) overlaps the Tasmanian World Heritage Area.

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 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

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,

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.

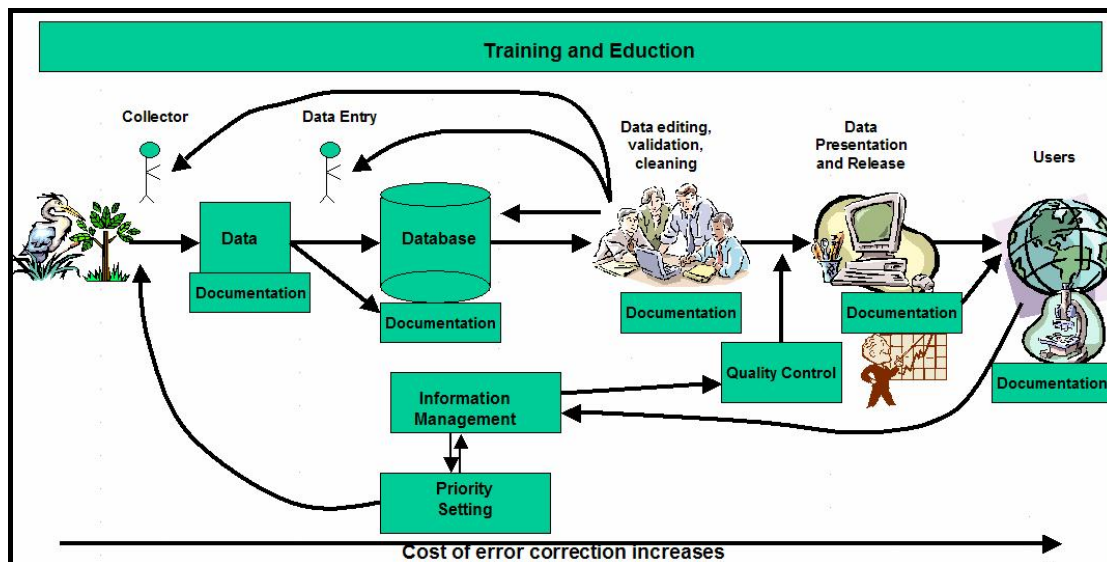


Fig. 3. Information Management Chain showing that the cost of error correction increases as one progresses along the chain. Good documentation, education and training are integral to all steps.

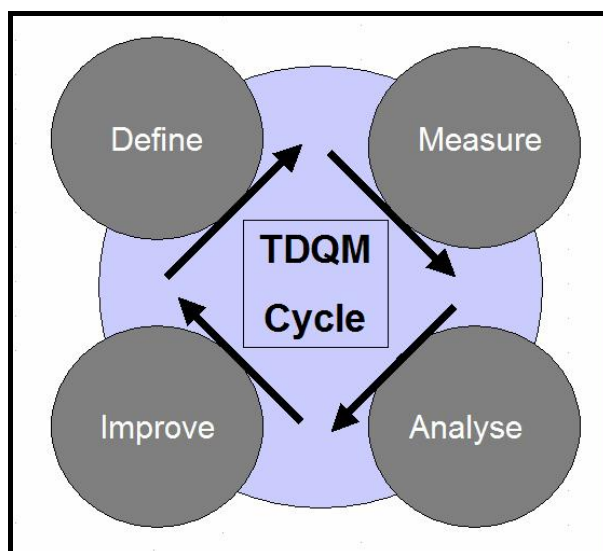


Fig. 4. *The Total Data Quality Management cycle showing the cyclical nature of the data management process (after Wang 1998).*

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 equally 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

² <http://www.tdwg.org/>

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.



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.

Data Cleaning

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
- 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).

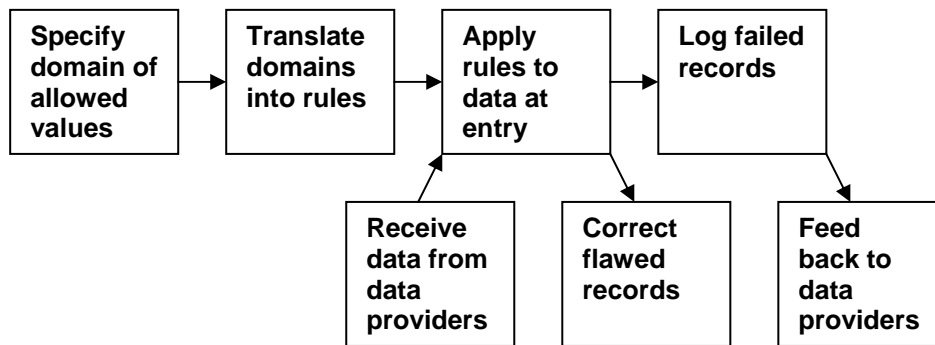


Fig. 5. *Use of edit controls (modified from Redman 2001).*

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 vouchered specimens are not

³ <http://www.webopedia.com/TERM/G/GUID.html>

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 hystrix* – 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 vouchers in all cases or areas.

Where vouchers are 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 root-mean-square deviation (RMSE) from zero to determine accuracy (Chrisman 1991). RMSE is not easy to apply to individual points, however, and

⁴ <http://www.ec-gis.org/inspire/>

⁵ <http://www.iso.ch/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020&ICS1=35>

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 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

⁶ <http://www.biogeomancer.org/>

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 increasingly 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 unskilled 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>

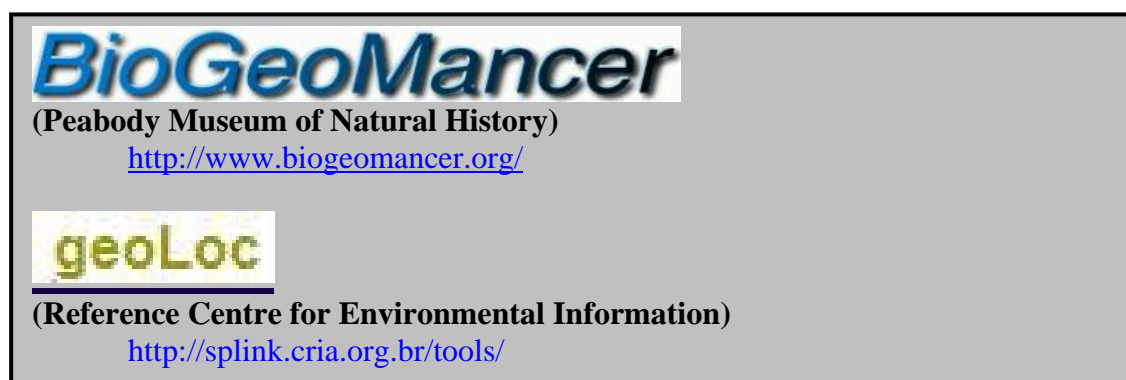
MaPSTeDI

Georeferencing in MaPSTeDI

<http://mapstedi.colorado.edu/geo-referencing.html>

A 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*.



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 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.



The data must be documented with sufficient detailed metadata to enable its use by third parties without reference to the originator of the data.

Search MaPSTeDI Collections

Taxon:

Common Name:

Locality:

State:

County:

Accuracy:

Institution:

Date from:

Date to:

☐ Abridged Record

☐ Complete Record

[Advanced Search](#)

[Basic Map](#)

Results per page:

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.

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 the 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 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.

⁷ <http://www.tdwg.org>;
<http://www.gbif.org/links/standards>



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 with incorrect data values in taxonomic

⁸ <http://www.cpbr.gov.au/avh/>

⁹ <http://specieslink.cria.org.br/>

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 tankervillae
 - 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 reporting techniques need to be developed and used, so that users of these data have similar abilities to correctly interpret and use the data.



Effective data quality programs help prevent embarrassment to the organisation and individuals – both internally and publicly.

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).

Acknowledgements

Many colleagues and organisations around the world have contributed to this paper in one way or another. Some directly, some by being involved in discussions with the author over a period of more than 30 years and some indirectly through published papers or just by making their information available to the world.

In particular, I would like to particularly make mention of the staff, both past and present, of CRIA (Centro de Referência em Informação Ambiental) in Campinas, Brazil and ERIN (Environmental Resources Information Network) in Canberra, Australia who have contributed ideas, tools, theories and a sounding board that have helped the author in formulating his ideas. Their discussion of error and accuracy in environmental information over the years and the pioneering work done by them, by CONABIO in Mexico, the University of Kansas, CSIRO in Australia, the University of Colorado, the Peabody Museum in Connecticut, and the University of California in Berkeley, as well as others too numerous to mention, has helped bring us to the stage we are today in species data quality management. I thank them for their ideas and constructive criticism. In addition, discussions with Town Peterson and others at the University of Kansas, Barry Chernoff at the Wesleyan University in Connecticut, Read Beaman at Yale University, John Wieczorek and Robert Hijmans at the University of California, Berkeley, Peter Shalk and others at ETI, in Amsterdam, Stan Blum at the Californian Academy and the staff of GBIF in Copenhagen have presented me with ideas and challengers that have led to some of the ideas expressed in this paper. Any errors, omissions or controversies are, however, the responsibility of the author.

I would like to also thank those who have supplied criticisms, comments and suggestions during the editing of this document, and in particular the following members of the GBIF Subcommittee for Digitisation of Natural History Collection Data: Anton Güntsch, Botanic Garden and Botanical Museum Berlin-Dahlem, Germany; Francisco Pando, Real Jardín Botánico, Madrid, Spain; Mervyn Mansell, USDA-Aphis, Pretoria, South Africa; A. Townsend Peterson, University of Kansas, USA; Tuuli Toivonen, University of Turku, Finland; Anna Wietzman, Smithsonian Institution, USA as well as Patricia Mergen, Belgian Biodiversity Information Facility, Belgium.

Larry Speers of GBIF was instrumental in the commissioning of the report, and in shepherding it through all its stages.

In conclusion I would like to thank the FAPESP/Biota project in Brazil with providing me with the opportunity and support to expand my ideas on data quality management during my stay in Brazil in 2003-2004 and the GBIF organisation for supporting and encouraging the production of this report.

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 Jatton, 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.rbgsyd.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.iconnect.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., Czaplewski, 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://mcmweb.er.usgs.gov/sdts/whatsdts.html> [Accessed 30 Jun. 2004].
- Van Sickle, 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 Charels, 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