PRINCIPLES AND METHODS OF DATA CLEANING

PRIMARY SPECIES AND SPECIES-OCCURRENCE DATA

Arthur D. Chapman[1]

Error qui non resistitur, approbatur

An error not resisted is approved.

(Ref. *Doct. & Stud.* c. 770).

Keywords:

Data Cleaning, Data Cleansing,

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

Chapman, A. D. 2005. *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data*, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen.

This paper was commissioned from Dr. 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
Helophilus pendulus (Linnaeus, 1758)

Contents

Data Cleaning

Data Cleaning is an essential part of the Information Management Chain as mentioned in the associated document, *Principles of Data Quality* (Chapman 2005a). As stressed there, error prevention is far superior to error detection and cleaning, as it is cheaper and more efficient to prevent errors than to try and find them and correct them later. No matter how efficient the process of data entry, errors will still occur and therefore data validation and correction cannot be ignored. Error detection, validation and cleaning do have key roles to play, especially with legacy data (e.g. museum and herbarium data collected over the last 300 years), and thus both error prevention and data cleaning should be incorporated in an organisation's data management policy.

One important product of data cleaning is the identification of the basic causes of the errors detected and using that information to improve the data entry process to prevent those errors from re-occurring.

This document will examine methods for preventing as well as detecting and cleaning errors in primary biological collections databases. It discusses guidelines, methodologies and tools that can assist museums and herbaria to follow best practice in digitising, documenting and validating information. But first, it will set out a set of simple principles that should be followed in any data cleaning exercises.

Definition: Data Cleaning

A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. The process may include format checks, completeness checks, reasonableness checks, limit checks, review of the data to identify outliers (geographic, statistical, temporal or environmental) or other errors, and assessment of data by subject area experts (e.g. taxonomic specialists). These processes usually result in flagging, documenting and subsequent checking and correction of suspect records. Validation checks may also involve checking for compliance against applicable standards, rules, and conventions.

The general framework for data cleaning (after Maletic & Marcus 2000) is:

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

There are a number of terms used by different people to refer largely to the same process. It is a matter of preference what one uses. Terms include:

- Error Checking;
- Error Detection;
- Data Validation;
- Data Cleaning;
- Data Cleansing;
- Data Scrubbing; and
- Error Correction.

I tend to use the term *Data Cleaning* to encompass three sub-processes, viz.

- Data checking and error detection;
- Data validation; and
- Error correction.

A fourth – improvement of the error prevention processes – could perhaps be added.

The Need for Data Cleaning

The need for data cleaning is centred around improving the quality of data to make them “fit for use” by users through reducing errors in the data and improving their documentation and presentation (see associated document on *Principles of Data Quality* – Chapman 2005a). Errors in data are common and are to be expected. Redman (1996) suggested that unless extraordinary efforts have been taken, that a field error rate of 1-5% should be expected. The usual view of errors and uncertainties is that they are bad, but a good understanding of errors and error propagation can lead to active quality control and managed improvement in the overall data quality (Burrough and McDonnell 1998). Errors in spatial position (geocoding) and in identification are two of the major causes of error in species-occurrence data and it is the cleaning of these errors that is covered in this paper. Correcting errors in data and eliminating bad records can be a time consuming and tedious process (Williams *et al.* 2002) but it cannot be ignored. It is important, however, that errors not just be deleted, but corrections documented and changes traced. As mentioned in the companion document on *Principles of Data Quality*, it is best to add corrections to the database while retaining the original data in a separate field or fields so that there is always the chance of going back to the original information.

Where are the Errors?

Primary species data encompass a whole range of data – from museum and herbarium data, through observational data (point-based, regional or area-based, and systematic or grid-based), to survey data, both systematic and other (Chapman 2005a). Because of the historical nature of many museum and herbarium collections (often termed legacy data); many records carry little geographic information other than a general description of the location where they were collected (Chapman and Milne 1998). With historical data, where geocodes coordinates are given they are often not very accurate (Chapman 1999) and have generally been added at a later date by those other than the collector (Chapman 1992). Many of these data have drawbacks when it comes to use for species’ distribution studies. Observational and survey data are also valuable records for many studies and the georeferencing information may be quite accurate, but because vouchered reference material is seldom retained, the taxonomic or nomenclatural information are generally less reliable than for documented museum collections. The georeferencing of survey and observational records may, however, still include errors or ambiguities, for example it may not be clear as to whether the geocode refers to the centre of the grid, or one corner in grid-based records.

Much of the data (both museum and observational) have been collected opportunistically rather than systematically (Chapman 1999, Williams *et al.* 2002) and this can result in large spatial biases – for example, collections that are highly correlated with road or river networks (Margules and Redhead 1995, Chapman 1999, Peterson *et al.* 2002, Lampe and Riede 2002). Museum and herbarium data and most observational data, generally only supply information on the presence of the entity at a particular time and says nothing about absences in any other place or time (Peterson *et al.* 1998). This restricts their use in some environmental models, but they remain the only complete collection of biological information covering the last 200+ years. The cost of replacing these data with new surveys would be prohibitive. It is not unusual for a single survey to exceed $1 million (Burbidge 1991). Further, because of their collection over time, they provide irreplaceable baseline data about biological diversity during a time when humans have had tremendous impact on such diversity (Chapman and Busby 1994). They are an essential resource in any effort to conserve the environment, as they provide the only fully documented record of the occurrence of species in areas that may have undergone habitat change due to clearing for agriculture, urbanisation, climate change, or been modified in some other way (Chapman 1999).

Preventing Errors

As stressed previously, the prevention of errors is preferable to their later correction, and new tools are being developed to assist institutions in preventing errors.

Tools are being developed to assist the process of adding georeferencing information to databased collections. Such tools include eGaz (Shattuck 1997), geoLoc (CRIA 2004a), BioGeomancer (Peabody Museum *n.dat*.), GEOLocate (Rios and Bart *n.dat*.) and the Georeferencing Calculator (Wieczorek (2001b). A project funded in 2005 by the Gordon and Betty Moore Foundation and involving worldwide collaboration is now bringing many of these tools together with the aim of making them available both as stand-alone open-source software tools and as Web Services. The need for further and more varied validation tools cannot, however, be denied. These tools will be discussed more fully later in this paper.

Tools are also being developed to assist in reducing error with taxonomic and nomenclatural data. There are two main causes of error with these data. They are inaccurate identifications or misidentifications (the taxonomy) and misspellings (the nomenclature). Tools to assist with the identification of taxa include improved taxonomies, floras and faunas (both regional and local), automated and computer-based keys to taxa, and digital imaging of type and other specimens. With the spelling of names, global, regional and taxonomic name-lists are being developed which allow for the development of authority files and database entry checklists that reduce error at data entry.

Perhaps the best way of preventing many errors is to properly design the database in the first instance. By implementing sound Relational Database philosophy and design any information that is frequently repeated such as species' names, localities and institutions, need only be entered once, and verified at the outset. Referential integrity then protects the accuracy of future entries.

Spatial Error

In determining the quality of species data from a spatial viewpoint, there are a number of issues that need to be examined. These include the identity of the specimen or observation – a misidentification may place the record outside the expected range for the taxon and thus appear to be a spatial error – errors in the geocoding (latitude and longitude), and spatial bias in the collection of the data. The use of spatial criteria may help in finding taxa that may be misidentified as will as misplaced geographically. The issue of spatial bias –very evident in herbarium and museum data (e.g. collections along roads) is more an issue for future collections, and future survey design rather than being related to the accuracy of individual plant or animal specimen records (Margules and Redhead 1995, Williams *et al*. 2002). Indeed – collection bias is more related to the totality of collections of an individual species, than it is to any one individual collection. In order to improve the overall spatial and taxonomic coverage of biological collections within an area, and hence reduce spatial bias, existing historical collection data can be used to determine the most ecologically valuable locations for future surveys for example, by using environmental criteria (climate etc.) as well as geographic criteria (Neldner *et al.* 1995).

Nomenclatural and Taxonomic Error

Names form the major key for accessing information in primary species databases. If the name is wrong, then access to the information by users will be difficult, if not impossible. In spite of having rules for biological nomenclature for around 100 years, the nomenclatural and taxonomic information in a database (the *Classification Domain* of Dalcin 2004) is often the most difficult in which to detect and clean errors. It is also the area that causes the most angst and loss of confidence amongst users in primary species databases. This is often due to ignorance amongst users of the need for taxonomic changes and nomenclatural changes, but is also partly due to taxonomists not fully documenting and explaining these changes to users, complications in the relationship between names and taxa, and confusion with taxonomic concepts that are often not well covered in primary species databases (Berendsohn 1997).

The easier of these errors to clean is the nomenclatural data – the misspellings. Lists of names (and synonyms) are the key tools for helping with this task. Many lists already exist for regions and/or taxonomic groups, and these are gradually being integrated into global lists (Froese and Bisby 2002). There are still many regions of the world and taxonomic groups, however that do not have reliable lists.

Taxonomic error – the inaccurate identification or misidentification of the collection is the most difficult of errors to detect and clean. Museums and herbaria have traditionally had a *determinavit* system in operation whereby experts working in taxonomic groups examine the specimens from time to time and determine their circumscription or identification. This is a proven method, but one that is time-consuming, and largely haphazard. There is unlikely to be any way around this, however, as automated computer identification is unlikely to be an option in the near or even long-term future. There are, however, many tools available to help with this process. They comprise both the traditional taxonomic publications with which we are all familiar and newer electronic tools. Traditional tools include publications such as taxonomic revisions, national and regional floras and faunas, and illustrated checklists. Newer tools include automated and computer-generated keys to taxa; interactive electronic publications with illustrations, descriptions, keys, and illustrated glossaries; character-based databases; imaging tools; scientific image databases that include images of types; systematic images of collections; and easily accessible on-line images (both scientifically verified and others).

Merging Databases

The merging of two or more databases will both identify errors (where there are differences between the two databases) and create new errors (i.e. duplicate records). Duplicate records should be flagged on merging so that they can be identified and excluded from analysis in cases where duplicate records may bias an analysis, etc., but should generally not be deleted. While appearing to be duplicates, in many cases the records in the two databases may include some information that is unique to each, so just deleting one of the duplicates (known as ‘Merge and Purge’ (Maletic and Marcus 2000)) is not always a good option as it can lead to valuable data loss.

An additional issue that may arise with merging databases is the mixing of data that are based on different criteria such as different taxonomic concepts, different assumptions or units of measurements and different quality control mechanisms. Such merging should always document the source of the individual data units so that data cleaning processes can be carried out on data from the different sources in different ways. Without doing this, it may make the database more difficult to clean effectively and to effectively document any changes.

Principles of Data Cleaning

Many of the principles of data cleaning overlap with general data quality principles covered in the associated document on *Principles of Data Quality* (Chapman 2005a). Key principles include:

Planning is Essential (Developing a Vision, Policy and Strategy)

Good planning is an essential part of a good data management policy. The Information Management Chain (figure 1) (Chapman 2005a), includes Data Cleaning as a central portion that needs to be incorporated into the organisation's data quality vision and policy. A strategy to implement data cleaning and validation into the organisation's culture will improve the overall quality of the organisation's data and improve its reputation with users and suppliers alike.

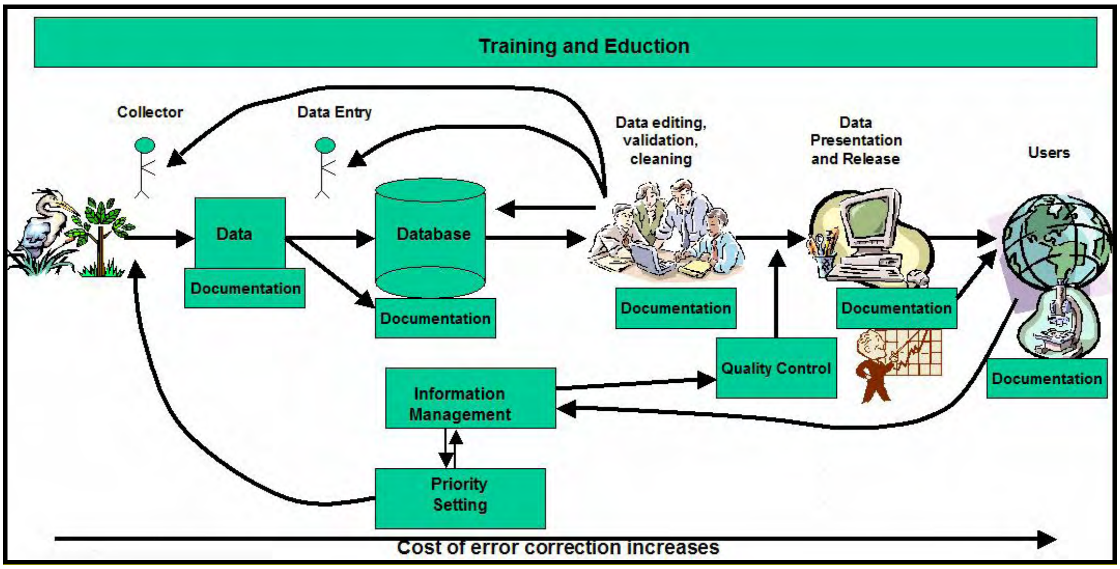


Fig. 1. *Information Management Chain showing that the cost of error correction increases as one moves along the chain. Education, Training and Documentation are integral to all steps (from Chapman 2005a).*

Organizing data improves efficiency

Organizing data prior to data checking, validation and correction can improve efficiency and considerably reduce the time and cost of data cleaning. For example, by sorting data on location, efficiency gains can be achieved through checking all records pertaining to one location at the same time, rather than going back and forth to key references. Similarly, by sorting records by collector and date, it is possible to detect errors where a record may be at an unlikely location for that collector on that day. Spelling errors in a variety of fields may also be found in this way.

Prevention is better than cure

As stressed previously (Chapman 2005a), it is far cheaper and more efficient to prevent an error, than to have to find it and correct it later. It is also important that when errors are detected, that feedback mechanisms ensure that the error doesn't occur again during data entry, or that there is a much lower likelihood of it re-occurring. Good database design will ensure that data entry is controlled so that entities such as taxon names, localities and persons are only entered once and verified at the time of entry. This can be done through use of drop-down menus or through keystroke identification of existing entries within a field.

Responsibility belongs to everyone (collector, custodian and user).

Responsibility for data cleaning belongs to all. The primary responsibility of the data-cleaning portion of the Information Management Chain (figure 1) obviously belongs to the data custodian – the person or organisation with principal responsibility for managing and storing the data. The collector, too, has responsibility and needs to respond to the custodian's questions when the custodian finds errors or ambiguities that may refer back to the original information supplied by the collector. These may relate to ambiguities on the label, errors in the date or location, etc. As will become obvious later in this document, the user also has a key responsibility to feed back to custodians information on any errors or omissions they may come across, including errors in the documentation associated with the data. It is often the user, when analysing or looking at the data in the context of other data, who will identify errors and outliers in the data that would otherwise go un-noticed. A single museum may have only a subset of the total available data (from one State or region for example), and it is only when the data are combined with data from other sources that errors may become obvious. Many of the tools elaborated in this document perform much better when looking at the totality of data for a species, collector or expedition than at subsets of them.

Partnerships improve Efficiency

Partnerships can be a very efficient method for managing data cleaning. As mentioned, the user is often the one who will be in the best position to identify errors in the data. If data custodians can develop partnerships with these key users then those errors won't be ignored. By developing partnerships, many data validation processes won't need to be duplicated, errors will more likely be documented and corrected, and new errors won't be incorporated by inadvertent "correction" of suspect records that are not in error. It is important to make these partnerships with users inside the organisation as well as outside as discussed in the associated paper on *Principles of Data Quality*.

Prioritisation reduces Duplication

As with organisation and sorting, prioritisation helps reduce costs and improves efficiency. It is often of value to concentrate on those records where extensive data can be cleaned at the lowest cost. For example, those that can be examined using batch processing or automated methods, before working on the more difficult records. By concentrating on those data that are of most value to users, there is also a greater likelihood of errors being detected and corrected. This improves client/supplier relationships and reputations, and provides greater incentive for both data suppliers and users to improve the quality of the data because it has an immediate use.

Setting of Targets and Performance Measures

Performance measures are a valuable addition to quality control procedures, and are used extensively with spatial metadata. They also help an organisation to manage their data cleaning processes. As well as providing users with information on the data and on their quality, such measures can be used by managers and curators to track those parts of the database that may need attention. Performance measures may include statistical checks on the data (for example, 95% of all records have an accuracy of less than 5,000 meters from their reported position), on the level of quality control (for example – 65% of all records have been checked by a qualified taxonomist within the previous 5 years; 90% have been checked by a qualified taxonomist within the previous 10 years), completeness (e.g. all 10-minute grid squares have been sampled) (Chapman 2005a).

Minimise duplication and reworking of data

Duplication is a major factor with data cleaning in most organisations. Many organisations carry out georeferencing at the same time as they database the record. As records are seldom sorted geographically, this means that the same or similar locations will be chased up a number of times. By carrying out the georeferencing of collections that only have textual location information and no coordinate information as a special operation, records from similar locations can then be sorted and located on the appropriate map-sheet or gazetteer. Some institutions also use the database itself to help reduce duplication by searching to see if the location has already been georeferenced (see under *Data Entry and Georeferencing*, below).

The documentation of validation procedures (preferably in a standardised format) is also important to reduce the reworking of data. For example, data quality checks carried out on data by a user may identify a number of suspect records. These records may then be checked and found to be valid records and genuine outliers. If this information is not documented in the record, further down the line, someone else may come along and carry out more data quality checks that again identify the same records as suspect. This person may then spend more valuable time rechecking the information and reworking the data. When designing databases, a field or fields should be included that indicates whether the data have been checked, by whom and when and with what result.

Experience in the business word has shown that the use of information chain management (see figure 1) can reduce duplication and re-working of data and lead to a reduction of error rates by up to 50% and a reduction in costs resulting from the use of poor data by up to two thirds (Redman 2001). This is largely due to efficiency gains through assigning clear responsibilities for data management and quality control, minimising bottlenecks and queue times, minimising duplication through different staff re-doing quality control checks, and improving the identification of better and improved methods of working (Chapman 2005a).

Feedback is a two-way street

Users of the data will inevitably carry out error detection, and it is important that they feedback the results to the custodians. As already mentioned, the user often has a far better chance of detecting certain error types through combining data from a range of sources, than does each individual data custodian working in isolation. It is essential that data custodians encourage feedback from users of their data, and implement the feedback that they receive (Chapman 2005a). Standard feedback mechanisms need to be developed, and procedures for receiving feedback agreed between the data custodians and the users. Data custodians also need to convey information on errors to the collectors and data suppliers where relevant. In this way there is a much higher likelihood that the incidence of future errors will be reduced and the overall data quality improved.

Education and Training improves techniques

Poor training, especially at the data collection and data entry stages of the Information Quality Chain, is the cause of a large proportion of the errors in primary species data. Data collectors need to be educated about the requirements of the data custodian and users of the data so that the right data are collected (i.e. all relevant parts and life stages), that the collections are well documented – i.e. the locality information is well recorded (for example – does 10 km NW of Town ‘y’ mean 10 km by road, or in a direct line), that standards are applied where relevant (e.g. that the same grid size is used for related surveys), and that the labels are clear and legible and preferably laid out in a consistent manner to make it easier for data entry operators.

The training of data entry operators is also important as identified in the MaPSTeDI georeferencing guidelines (University of Colorado Regents 2003a). Good training of data entry operators can reduce the error associated with data entry considerably, reduce data entry costs and improve overall data quality.

Accountability, Transparency and Audit-ability

Accountability, transparency and audit-ability are essential elements of data cleaning. Haphazard and unplanned data cleaning exercises are very inefficient and generally unproductive. Within data quality policies and strategies – clear lines of accountability for data cleaning need to be established. To improve the “fitness for use” of the data and thus their quality, data cleaning processes need to be transparent and well documented with a good audit trail to reduce duplication and to ensure that once corrected, errors never re-occur.

Documentation

Documentation is the key to good data quality. Without good documentation, it is difficult for users to determine the fitness for use of the data and difficult for custodians to know what and by whom data quality checks have been carried out. Documentation is generally of two types and provision for them should be built into the database design. The first is tied to each record and records what data checks have been done and what changes have been made and by whom. The second is the metadata that records information at the dataset level. Both are important, and without them, good data quality is compromised.

Methods of Data Cleaning

Introduction

Museums and herbaria throughout the world are beginning to database their collections at increasing rates, and are starting to make at least some of that information available via the Internet. The rate of databasing of collections has increased in recent years with the development of tools and methodologies that can assist in the process, and increased publication since the creation of the Global Biodiversity Information Facility (GBIF) with its aim to "make the world's primary data on biodiversity freely and universally available via the Internet" (GBIF 2003a).

As well as good practices (see associated document – *Principles of Data Quality* and *Principles of Data Cleaning*, this document), there is a need for useful and powerful tools that automate, or greatly assist in the data cleaning process. Automated methods can only be part of the procedure and there is a continuing need for the development of new tools to assist this process, and for their use to be integrated into best practice routines. Manual cleaning of data is laborious and time consuming, and is in itself prone to errors (Maletic and Marcus 2000), but it will continue to be important with primary species-occurrence databases. Where possible, it should only be carried out as a last resort, for small data sets, or where other checks have left just a few errors that cannot be checked any other way.

Some of the techniques that have been developed include the use of climate models to identify outliers in climate space (Chapman 1992, 1999, Chapman and Busby 1994), in geographic space (CRIA 2004b, Hijmans *et al*. 2005, Marino *et al.* in prep.), the use of automated georeferencing tools (Beaman 2002, Wieczorek and Beaman 2002) and many others. Most collection institutions do not have a high level of expertise in data management techniques or in Geographic Information Systems (GIS). What is needed in these institutions is a simple, inexpensive set of tools to both assist in the input of data and information, including geocoding information, and similar simple and inexpensive tools for data validation that can be used without the necessary incorporation of expensive GIS software. Some tools have been developed to assist with data entry – tools such as Biota (Colwell 2002), BRAHMS (University of Oxford 2004), Specify (University of Kansas 2003a), BioLink (Shattuck and Fitzsimmons 2000), Biótica (Conabio 2002), and others that provide database management and associated data entry (Podolsky 1996, Berendsohn *et al.* 2003); and eGaz (Shattuck 1997), geoLoc (CRIA 2004a, Marino *et al.* in prep.), GEOLocate (Rios and Bart *n.dat.*) and BioGeoMancer (Peabody Museum *n.dat.*), that assist in the georeferencing of collections. There are also a number of documented guidelines available on the Internet that can assist institutions in setting up and managing their databasing programs. Examples include the MaNIS Georeferencing Guidelines (Wieczorek 2001a), MaPSTeDI Georeferencing Guidelines (University of Colorado Regents 2003a) and HISPID (Conn 1996, 2000).

There are many methods and techniques that can aid in the cleaning of errors in primary species and species-occurrence databases. They range from methods that have been operating in museums and herbaria for hundreds of years, to automated methods that are still largely untested. This paper looks in detail at a range of methods for cleaning species databases, and where possible, provides examples. It is by no means a comprehensive list as many institutions have developed their own techniques and methodologies.

Because of the very nature of natural history collections, it is not possible that all geocode information be highly precise, or that there is a consistent level of precision within a database. Data with a very low precision, however, are not necessarily of low quality. Quality only comes into being once the data are being used and is not a character of the data *per se* (see discussion in associated paper on *Principles of Data Quality* - Chapman 2005a). Quality is merely a factor of fitness for use or potential use and is a relative term. What is important is for users of the data to be able to determine from the data themselves, if the data are likely to be fit for the required application. The level of accuracy of each given geocode should therefore be recorded within the database. I prefer this to be in non-categorical form, recorded in meters, however many databases have developed categorical codes for this purpose. When this information is available, a user can request, for example, only those data that are better than a certain metric value – e.g. better than 5,000 meters (see example using codes for extracting data at University of Colorado Regents 2003b). There are a number of ways of determining accuracy of geocoded records. The point-radius method (Wieczorek *et al.* 2004) is, I believe, the easiest and most practical method, and is one previously recommended for use in Australia (Chapman and Busby 1994). It is also important that automated georeferencing tools include calculated accuracy as a field in the output. The geoLoc (CRIA 2004a, Marino *et al.* in prep.) and BioGeomancer (Peabody Museum *n.dat.*) tools, which are still under development, include this feature.

Over time, it is hoped that species collection data resources will improve as institutions move to more precise instrumentation (such as GPS) for recording the location of new records and as historic records are corrected and improved. It is also important that collectors make the best possible use of the tools available to them and not just use a GPS to record data to 1 arc minute resolution because of historical reasons - as that is the finest they have recorded information prior to using a GPS. If this is done, then they should make sure that the appropriate accuracy is added to the database otherwise it may be assumed that as a GPS was used, the accuracy is 10 meters as opposed to the 2000 meters of reality. Error prevention is preferable to error detection, but the importance of error detection cannot be under stressed, as error prevention alone can never be guaranteed to prevent all possible errors.

Taxonomic and Nomenclatural Data

Names, whether they are scientific binomials or common names, provide the first point of entry to most species and species-occurrence databases. Errors in names may arise in a number of ways: the identification may be wrong, the name may be misspelt, or the format may be wrong (or not what is expected by the user). The first of these is not easy to check or rectify without tedious effort, and requires the services of a taxonomic expert. The others though, are more easily catered for with good database design and methods that assist with data entry so that these errors do not occur or are minimised.

A. Identification certainty

Traditionally, museums and herbaria have had an identification or "*determinavit*" system in operation whereby experts working in taxonomic groups from time to time examine the specimens and determine their identifications. This may be done as part of a larger revisionary study, or by an expert visiting another institution and, while there, checks the collections. This is a proven method, but one that is time-consuming, and largely haphazard. There is unlikely to be any way around this, however, as automated computer identification is unlikely to be an option in the near or even long-term future.

i. Database design

One option may be the incorporation of a field in databases that provides some indication of the certainty of the identification when made. There are a number of ways that this could be done, and it perhaps needs some discussion to develop a standard methodology. This would be a code field, and may be along the lines of:

- identified by World expert in the taxa with high certainty
- identified by World expert in the taxa with reasonable certainty
- identified by World expert in the taxa with some doubts
- identified by regional expert in the taxa with high certainty
- identified by regional expert in the taxa with reasonable certainty
- identified by regional expert in the taxa with some doubts
- identified by non-expert in the taxa with high certainty
- identified by non-expert in the taxa with reasonable certainty
- identified by non-expert in the taxa with some doubt
- identified by collector with high certainty
- identified by collector with reasonable certainty
- identified by collector with some doubt

How one might rank these would be open to some discussion, and likewise whether these were the best categories or not. There are apparently some institutions that already do have a field of this nature. The HISPID Standard Version 4 (Conn 2000) does include a simplified version – the Verification Level Flag with five codes (Table 1).

Many institutions also already have a form of certainty recording with the use of terms such as: "aff.", "cf.", "*s. lat*.", "*s. str*.", "?". Although some of these (aff., cf.) have strict definitions, their use by individuals can vary considerably. The use of *sensu stricto* and *senso lato* imply variations in the taxonomic concept rather than levels of certainty, although not always used in that way.

0	The name of the record has not been checked by any authority
1	The name of the record determined by comparison with other named plants
2	The name of the record determined by a taxonomist or by other competent persons using herbarium and/or library and/or documented living material
3	The name of the plant determined by taxonomist engaged in systematic revision of the group
4	The record is part of type gathering or propagated from type material by asexual methods

***Table 1.** Verification Level Flag in HISPID (Conn 2000).*

As an alternative, where names are derived from other than taxonomic expertise, one could list the source of the names used (after Wiley 1981):

- descriptions of new taxa
- taxonomic revisions
- classifications
- taxonomic keys
- faunistic or floristic studies
- atlases
- catalogues
- checklists
- handbooks
- taxonomic scholarship/rules of nomenclature
- phylogenetic analysis

ii. Data entry

As data are being entered into the database, checks can be made as to whether the name has been checked by an expert or not, and if any of the above fields are present, that they have been entered. If such fields are used, they should be entered through use of a check-list or authority file that restricts the available options and thus reduces the chance of errors being added.

iii. Error checking

Geocode checking methods (see under *Spatial Data*, below) can also help identify misidentifications or inaccurate identifications through the detection of outliers in geographic or environmental space. Although generally an outlier found through geocode checking will be an error in either the latitude or longitude, occasionally it indicates that the specimen has been misidentified as the taxon being studied and thus falls outside the normal climate, environmental or geographic range of the taxon. See below for a more detailed discussion on techniques for identifying geographic outliers.

The main method for detecting whether a collection is accurately identified or not, though, is for experts to check the identification by examining the specimen or voucher collections where they exist. Geocode outlier detection methods cannot determine if a collection is accurately identified or not, but may help identify priority collections for expert taxonomic checking. With observational data, experts may be able to determine, on personal knowledge, if the taxon is a likely record for the area (e.g. Birds Australia 2004); but generally it is difficult to identify an inaccurate identification of an observational record where there are no voucher specimens. Many institutions may flag doubtful or suspect records and then it is up to the user to decide if they are suitable for their use or not.

B. Spelling of names

This paper does not attempt to cover all the possible kinds of names that may be entered into a primary species database. For example, hybrids and cultivars in plant databases, synonyms of various kinds, and taxonomic concepts all have specific issues and the checking of these can be problematic. Examples of how such names may be treated can be found in the various International Codes of Nomenclature, as well as in TDWG Standards such as HISPID (Conn 1996, 2000) and Plant Names in Botanical Databases (Bisby 1994).

a. Scientific names

The correct spelling of a scientific name is generally governed by one of the various relevant nomenclature Codes. However, errors can still occur through typing errors, ambiguities in the Code, etc. The easiest method to ensure such errors are kept to a minimum is to use an ‘Authority File” during input of data. Most databases can be set up to incorporate either an unchangeable authority file, or an authority file that can be updated during input.

i. Database design

One of the keys to being able to maintain good data quality with taxon names is to split the name into individual fields (i.e. atomise the data) rather than maintain them all in one field (e.g. genus, species, infraspecific rank, infraspecific name, author and certainty). Maintaining these in one field reduces the opportunities for good data validation and error detection, and can lead to a quite considerable increase in the opportunity for errors. For example, by separating the genus and species parts of the name, each genus name only needs to be added once into a relational database (through an authority file or pick-list), thus reducing opportunities for typographic errors and misspellings.

Databases that include all parts of the name in one field can make it very difficult to maintain quality or to combine with other databases. It introduces another range of possible errors and is not recommended. Some databases use both – a combined field and atomised fields, but this again provides opportunity for added error if these are not automatically generated, and if one is updated and the other not. Automatically generated fields eliminate the danger of this.

Two issues that need to be considered with atomised data are the incorporation of data in a database where the data is in one field (for example the importing of lists of plants or animals), and the need to present a concatenated view of data from an atomised database, for example on a web site or in a publication.

With the first of these – the parsing of data from a concatenated field into individual fields is generally not as simple a process as it might appear. It is not only an issue with names, of course, as the same problems arise with references and locality information as discussed below. As far as I am aware, no simple tools for doing the parsing exist, however many museums and herbaria have done this for their own institutions, and thus algorithms may be available from these institutions.

With the second, the requirement to present a concatenated view on the output side for presentation on the Web or in reports could either be carried out using an additional (generated) field within the database that concatenates the various fields, or done on the fly during extraction. This is an issue that should be considered when designing the database and its reporting mechanisms. They are issues that the taxonomic community may need to discuss further with the aim of developing simple tools or methodologies.

ii. Authority files

Authority files exist for a number of taxonomic groups, and are being developed by a range of agencies. Reliable authority files are available for many higher taxa (Families, Orders, and Genera), and these can be used to ensure data integrity in these fields. It is unlikely that a detailed authority file for all taxa, especially to the species level and below, will be produced in the near future, however, existing authority files (e.g. IPNI 1999, Froese and Bisby 2004) can be used as a beginning. If authority files are available, then the databases can be set up in such a way that new names can be added to them. For example, assume a database has an authority file with a pull down list, or fills in the field as one types (for example as happens in an EXCEL spreadsheet if one starts to type a name in a field where that name may already be in an earlier row).

1. Use the pull down list to search for the name
2. It is not there
3. Click on the button – "New name"
4. Add the New Name
5. The database may come back and say "This name is similar to <name>" do you want to continue?
6. Yes
7. The name is added to the list, and the next time you wish to add a name, that name will now appear in the pull-down list.

In this way, you are gradually adding to and improving the authority file.

As an extra check, these names may then go into a secondary list that a supervisor can verify and either approve or discard. Depending on the level of sophistication of the database, the list may include synonyms and if you begin to type in a name, it may ask you if you really wish to add this name as it is listed in the authority file as a synonym of <name>.

It is recommended that Authority files be used wherever possible. A good start is the Species2000 & ITIS Catalogue of Life (Froese and Bisby 2004), available on CD as an annual checklist. The format of this document is being improved for future editions to make it easier to incorporate into databases. The checklist is also available electronically for checking individual taxa and is in addition to a regularly updated checklist, which is also available, on-line. Also, a number of names databases exist or are being developed and these can form the basis of a names authority file. Some examples include,

Global lists such as:

- Species2000 & ITIS Catalogue of Life (Froese and Bisby 2002),
- Ecat (GBIF 2003b),
- International Plant Name Index (IPNI 1999);
- Global Plant Checklist (IOPI 2003).

Regional lists such as:

- Integrated Taxonomic Information System (Ruggiero 2001);
- Australian Plant Name Index (Chapman 1991, ANBG 2003);
- Proyeto Anthos – Sistema de información sobre los plantas de España (Fundación Biodiversidad 2005)
- Australian Faunal Directory (ABRS 2004);
- Med Checklist (Greuter *et al.* 1984-1989).

Taxonomic lists such as:

- ILDIS World Database of Legumes (Bisby *et al.* 2002);
- Fishbase (Froese and Pauly 2004);
- World Spider Catalog (Platnik 2004);

- Many others.

Where authority files are imported from an external source such as one of those above, then the Source-Id should be recorded in the database so that changes that are made between editions of the authority source can be easily incorporated into the database, and the database updated. Hopefully, before long this may become easier through the use of Globally Unique Identifiers (GUIDs)[1].

iii. Duplicate entries

Even when designing a database from scratch and trying to normalise it as much as possible for example by using authority tables, the issue of duplicate records cannot be avoided, and especially when importing data from secondary sources (e.g. names or references). To remove (or flag) such duplicates a special interface may be needed. The interface should be capable of identifying potential duplicates using special algorithms. The data entry operator (or curator, expert, etc.) will then have to decide from the list of potential duplicates the set of records identified as real duplicates and the records that should be retained. The systems will then discard and archive, or flag the superfluous records while keeping the referential integrity of the system. Generic software could be implemented to rectify this, but as with the parsing software, does not appear to exist at the moment. Biodiversity database designers should be aware of the problem and consider designing a generic software tools for these tasks.

iv. Error checking

It is possible to carry out some automated checks on names. Although complete lists of species names do not exist, lists of family names, and generic names (e.g. IAPT 1997, Farr and Zijlstra *n.dat.*) are much more complete, especially for some taxa. Checks against these fields could be carried out against such lists. With species epithets (the second part of the binomial) there are a number of tests that can be conducted. For example, looking for names within the same genus that have a high degree of similarity – names with one character out of place or with a character or characters missing, etc. The CRIA Data Cleaning system (CRIA 2005) carries out many of these tests on distributed data obtained through speciesLink (CRIA 2002). Best practice in this case would be for automated detection, but not automated correction. Other possible checks include (modified from English 1999, Dalcin 2004):

Missing Data Values **–** This involves searching for empty fields where values should occur. For example, in a botanical database if an infraspecific name is cited, then a value should be present in the corresponding infraspecific rank field; or if a species name is present, then a corresponding generic name should also be present.

Incorrect Data Values **–** This involves searching for typographic errors, transposition of key strokes, data entered in the wrong place (e.g. a species epithet in the generic name field), and data values forced into a field that requires a value (i.e. is a mandatory field), but for which the data entry operator doesn’t know the value so adds a dummy value. There are a number of ways of checking for some of these errors – for example, using Soundex, (Roughton and Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), or Skeleton-Key (Pollock and Zamora 1984). Each of these methods uses slightly different algorithms for detecting similarity. A recent test of a number of methods (including those mentioned) using species names and a number of datasets (Dalcin 2004) showed that the Skeleton-Key method produced the highest proportion of true errors to false errors in the datasets tested. An on-line example of using these methods can be seen on the CRIA site in Brazil (CRIA 2005). These are further explained below.

Nonatomic Data Values **–** This involves searching for fields where more than one fact is entered. For example “subsp. bicostata” in the infraspecies field where this should be split into two fields.

[1] http://www.webopedia.com/TERM/G/GUID.html

Depending on database design (see above) this may not be an error. Nonatomic data values occur in many databases and are difficult to remove. An essential first step is that such values indicate that the database probably needs a new field created. Some nonatomic data can then be split into the relevant fields using automated methods, but more often than not, many are left that can only be fixed manually under the control of an expert.

Domain Schizophrenia – This involves searching for fields used for purposes for which they may not have been intended. This often happens where a certainty field has not been included in the database and question marks, uncertainties such as cf., aff. are added in the same field as the species epithet, or comments added (Table 2). The nature of this 'error' may also depend on database design.

Family	Genus	Species
Myrtaceae	Eucalyptus	globulus?
Myrtaceae	Eucalyptus	? globulus
Myrtaceae	Eucalyptus	aff. globulus
Myrtaceae	Eucalyptus	sp. nov.
Myrtaceae	Eucalyptus	?
Myrtaceae	Eucalyptus	sp. 1
Myrtaceae	Eucalyptus	To be determined

***Table 2**. Examples of Domain schizophrenia (from Chapman 2005a).*

Duplicate Occurrences – This involves searching for names that may refer to the same real world value. There are two main types of duplicates that can occur here – the first is an error due to misspellings, and the second is where there is more than one valid alternate name such as with the International Code of Botanical Nomenclature (2000) which allows for alternate Family names (e.g. Brassicaceae/Cruciferae, Lamiaceae/Labiatae). The latter can be handled by either choosing one of the valid alternatives for the database, or using linked synonyms depending on the policy of the institution. Similar issues may also occur where alternate classifications have been followed at higher taxonomic ranks, or even at the genus level where a species may validly occur in more than one genus depending on whose classification is followed (*Eucalyptus*/*Corymbia*; Albatross species in the Southern Hemisphere, small wild cat species, and many more).

Inconsistent Data Values – This occurs where two related databases do not use the same names lists, and when combined (or compared) show inconsistencies. For example, this may occur at botanic gardens between the Living Collection and the Herbarium; when merging databases of two specialists; or in museums between the collection database and the images database. Correcting involves checking one database against the other to identify the inconsistencies.

Dalcin (2004) conducted a number of detailed experiments on methods of checking for spelling errors in scientific names and developed a set of tools to check for phonetic similarity. I have not used or tested these tools, but details on the methods and the results and comparative tests between methods can be obtained from Dalcin (2004) pp. 92-148. Also, CRIA, in Brazil have developed name-checking routines along similar lines (CRIA 2005) and these are expanded in the methods section below.

b. Common names

There are no hard and fast rules for 'common' or vernacular names, be they in Portuguese, Spanish, English, Hindi, various other languages, or regionally-based indigenous names. Often what are called 'common' names are in reality colloquial names (especially in botany) and may have just been coined from a translation of the Latin scientific name. In some groups, for example birds (see Christidis and Boles 1994) and fish (Froese and Pauly 2004), agreed conventions and recommended

English names have been developed. In many groups the same taxon may have many common names, which are often region-, language-, or people-specific. An example is the species *Echium plantagineum* which is known variously as ‘Paterson’s Curse’ in one Australia State and ‘Salvation Jane’ in another and with other names (e.g. Viper’s Bugloss, Salvation Echium) in other languages and countries. Conversely, the same common name may be applied to multiple taxa, sometimes in different regions, but sometimes even in the same region.

It is just about impossible to standardise common names, even across one language except perhaps for some small groups. But does it make any sense to try and do this (Weber 1995)? True common names are names that have developed and evolved over time, and the purpose of having them is so people can communicate. What I am suggesting here is that common names not be standardised, but that when placed in a database it is done in a standard way and their source documented. Many users of primary species occurrence data want to access data through the use of common names, so there is value in having them in our databases if we want to make our data of the most use to the largest possible audience. By adopting standard methods for recoding common names, be it one per species or hundreds - in one language or in many – and documenting the source of each name, we can make searching and thus information retrieval that much more efficient and useful.

There are many difficulties in including common names in species databases. These include:

- names in non-Latin languages that require the use of Unicode within the database for storage. Problems may occur:
    - where databases attempt to store the names phonetically using just the Latin alphabet,
    - where people are not able to display the characters properly on their screen or in print,
    - in carrying out searches where users have only a “Latin” keyboard,
    - with data entry where names are a mix of Latin and non-Latin,
- the need to store information on the language of the name, especially where names of mixed language are included,
- the need to store information on regional factors – the area for which the name may be relevant, the language dialect, etc.
- the need to store information such as the references to the source of the name.

It is not any easy task to do properly and particularly in a way that increases usefulness while reducing error. If it is decided to include such names, the following may help in providing some degree of standardisation.

i. Data entry

When databasing common names, it is recommended that some form of consistency in construction be followed. It is probably most important that each individual database be internally consistent. The development of regional or national standards is recommended where possible. There are too many languages and regional variations to attempt to develop a standard for all languages and all taxa, although some of the concepts proposed here could form the basis for a standard in languages other than those covered.

For English and Spanish common names, I recommend that a similar convention to that developed for use in Environment Australia (Chapman *et al.* 2002, Chapman 2004) be followed. These guidelines were developed to support consistency throughout the organisation’s many databases. These conventions include beginning each word in the name with an initial capital.

Sunset Frog

With generalised or grouped names a hyphen is recommended. The word following the hyphen is generally not capitalised, except for birds where the word following the hyphen is capitalised if it is a member of a larger group as recommended by Christidis and Boles (1994).

> Yellow Spider-orchid
> Double-eyed Fig-Parrot ('Parrot' has an initial capital as it is a member of the Parrot group).

Portuguese common names are generally given all in lower case, usually with hyphens between all words if a noun, or separated by a space if a noun and adjective. It is recommend that for Portuguese common names, either this convention be followed or be modified to conform to the English and Spanish examples.

> mama-de-cadela,
> fruta-de-cera
> cedro vermelho

There is some disagreement and confusion as to whether apostrophes should be used with common names. For geographic names, there is a growing tendency to ignore all apostrophes (e.g. Smiths Creek, Lake Oneill), and it is now accepted practice in Australia (ICSM 2001, Geographic Names Board 2003 Art. 6.1). I recommend that a similar convention could be adopted with common names, although there is no requirement at present to do so.

Where names are added in more than one language and/or vary between regions, dialects or native peoples, then the language and the regional information should be included in a way that it can be attached to the name. This is best done in a relational database by linking to the additional regional and language fields, etc. In some databases, where there is only a language difference, the language is often appended to the name in brackets, but although this may appear to be a simple solution in the beginning, it usually becomes more complicated over time and often becomes unworkable. It is best to design the database to cater for these issues in the beginning rather than have to add flexibility at a later date.

If names in non-Latin alphabets are to be added to the database, then the database should be designed to allow for the inclusion of the UNICODE character sets.

ii. Error checking

As common names are generally tied to the scientific name, checks can be carried out from time to time to check for consistency within the database. This can be a tedious procedure, but only need be carried out at irregular intervals. Checks can be done by extracting all unique occurrences and checking for inconsistencies, e.g. missing hyphens.

Again, programs such as Soundex, (Roughton and Tyckoson 1985), Phonix (Pfeiffer *et al.* 1996), or Skeleton-Key (Pollock and Zamora 1984) could be used to search for typographic errors such as transposition of characters as mentioned above for *scientific names* (see Dalcin 2004).

c. Infraspecific rank

The use of an infraspecific rank field(s) is a more significant in databases of plants than in databases of animals. Animal taxonomists general only use one rank below species – that of subspecies (and even this is used with decreasing frequency), with the name treated as a trinomial:

> *Stipiturus malachurus parimeda.*

Historically, however, some animal taxonomists did use ranks other than subspecies, and databases may need to cater for these. If so, then the comments made for plant databases below, will also apply to databases of animal names.

i. Database design

As mentioned elsewhere, there are major data quality advantages in keeping the infraspecific rank separate from the infraspecific name. This allows for simple checking of the infraspecific rank field, and makes checking of the infraspecific name field easier as well. The rank and the name should never be included as "content" within the one field.

One issue that should be considered with atomised databases is the need in some cases to concatenate these fields for display on the web, etc. This can generally be automated, but consideration to how this may be done (whether in the database as an additional generated field or on the fly) should be considered when designing the database and its output forms.

ii. Data entry

With plants (and historically animals), there are several levels below species that may be used. These infraspecific ranks are most frequently ***subspecies, variety, subvariety, forma*** and ***subforma*** (the Botanical Code does not preclude inserting additional ranks, so it is possible that other ranks may exist in datasets). Subvariety and subforma are seldom used, but do need to be catered for in plant databases. Again, a pick-list should be set up with a limited number of choices. If this is not done, then errors begin to creep in, and you will invariably see subspecies given as: subspecies, subsp., ssp., subspp., etc. This can then be a nightmare for anyone trying to extract data, or to carry out error checking. It is better to restrict the options at the time of input, than have to cater for a full range at the time of data extraction, or attempt data cleaning to enforce consistency at a later date. It is recommended that the following be used:

subsp.	subspecies
var.	variety
subvar.	subvariety
f.	form/forma
subf.	subform
cv.	cultivar (often treated in databases as another rank, but see separate comments below)

In collection databases, the inclusion of a hierarchy is not necessary where more than one level may exist, because this just adds an extra layer of confusion and under the International Code for Botanical Nomenclature (2000) the hierarchy is unnecessary to unambiguously define the taxon. If the hierarchy is included, it must be possible to extract only that which is necessary to unambiguously define the taxon.

> *Leucochrysum albicans* subsp. *albicans* var. *tricolor* (= *Leucochrysum albicans* var. *tricolor*).

iii. Error checking

If the database has been designed well and a checklist of values used, then there is less need for further error checking. Where this is not the case, however, checks should be carried out to ensure that only the limited subset of allowed values occurs. One check that should be done however is for missing values.

d. Cultivars and Hybrids

Cultivars and hybrids occur in many plant databases and are often not handled well. Cultivars are subject to their own Code of Nomenclature (Brickell *et al.* 2004). In many plant species databases they are treated as just another infraspecific rank ("cv.") and in some databases this may be quite acceptable. Hybrids are much more difficult to handle than most other groups. They may be given a binomial name and can then be treated as any other taxon of the same rank (preceded by an "**X**" (multiplication sign) to denote a hybrid), or they may be treated as a formula (a cross between two, or more taxa which may even be at different ranks) indicated with taxonomic names separated by multiplication signs.

i. Database design

I recommend that anyone looking at setting up a database of plants that may include hybrids or cultivars consult the HISPID standard (Conn 1996, 2000) where hybrids are treated as part of the Record Identification Group. However they are handled, it is good practice to include a field that states that the name belongs to a cultivar or hybrid, etc. In this way they can be extracted separately and treated differently (for formatting, concatenation, etc.) on extraction, and for error checking.

ii. Error checking

Checking of errors for hybrids and cultivars is a difficult task if the database has not been set up to cater for it. One suggestion for checking may be to treat them as a group, i.e. extract all hybrids and sort them alphabetically by species depending on how they are stored in the database. This is much easier to do where a separate field is included that identifies hybrid records as such. One key error that is likely to occur is inconsistency with the use of the '**X**' sign. Some databases may not allow for a multiplication sign and it is commonly replaced by an 'x' or 'X' sometimes with a space before the name and sometimes not. These sorts of consistencies can easily be checked. I know of no really good system for checking errors in hybrid names.

e. Unpublished Names

i. Data entry

Not all records placed in a primary species databases are going to belong to a validly published taxon. To be able to retrieve these records from the database it is necessary to provide a 'temporary' name for that collection. If unpublished names can be incorporated into a database in a standard format, it makes it a lot easier to keep track of them, and to be able to retrieve them at a later date. It is also better, and less confusing than adding unpublished names that are binomials that look like published names, with or without the tag such as "nomen novum", "nom. nov." and "ms". Too often the "ms" or "nom. nov." is left off and users can spend a lot of time looking for the publication and reference information for the unpublished name. By using a formula it is obvious to all that it is an unpublished name.

In the 1980s in Australia, botanists agreed on a formula (Croft 1989, Conn 1996, 2000) for use with unpublished names. This was to avoid confusion arising through the use of such things as "*Verticordia* sp.1", "*Verticordia* sp.2" etc. Once databases begin to be combined, for example through the Australian Virtual Herbarium (CHAH 2002), *species*Link (CRIA 2002), Biological Collection Access Service for Europe (BioCASE 2003), the Mammal Networked Information System (MaNIS 2001), the GBIF Portal (GBIF 2004) and many others, names like these can cause even more confusion as there is no guarantee that what was called "sp.1" in one institution is identical to "sp.1" in a second. One way to keep these databases clean and consistent, and enable the smooth transfer of data from one to another, is through the use of a formula similar to that adopted by Australian botanical community.

The agreed formula is in the form of:

> "<Genus> sp. <colloquial name or description> (<Voucher>)":
>
> *Prostanthera* sp. Somersbey (B.J.Conn 4024)

Later, when the taxon is formally described and named, the formula-name can be treated as a synonym, just like any other synonym.

The use of such a formula makes a database more complicated than it may otherwise be, because instead of the species field only ever having one word; to cater for the formula it now requires inclusion of a sentence. The use of "sp. 1" "nom. nov." etc. as is often used have the same problem, and this method leaves less room for ambiguity. The use of formulae like these can cause difficulties with concatenation (for presentation on the web, etc.), however experience with the use of this methodology in Australia (for example, see the use with the SPRAT database of the Australian Department of the Environment (DEH 2005b)) has proved to work well. In all other ways, however, the formula is treated as a 'species" epithet, albeit with spaces and brackets, etc.

Because of the need to use unpublished names, for example in legal lists of threatened species (see for example, DEH 2005a), it is essential that there is a consistent system of naming or tagging these taxa for use in non-taxonomic publications, for example in legislative instruments. By using a formula like that suggested here, there is little danger of accidentally publishing a nomen nudum by mistake.

It is recommended that museum and herbaria adopt a similar system for use in their databases.

ii. Error checking

The most common error that occurs with a formula name such as suggested here is that of misspelling. Because the formula usually includes several words, it is often easy to make a mistake with citation of the voucher, etc. The easiest way in which to check such names is to sort each within a genus (they should be the only names in the species or infraspecies fields with more than one word) and examine them for similarities. This should not be too onerous a task as there is unlikely to be a huge number within any one genus. Similar techniques such as Soundex, mentioned above, could also be used.

f. Author names

The authors of species names may be included in some specimen databases, but more often than not, their inclusion can lead to error as they are seldom thoroughly checked before inclusion. They are only really necessary where the same name may have inadvertently have been given to two different taxa (homonyms) within the same genus or where the database attempts to include taxonomic concepts (Berendsohn 1997). The inclusion of the author's name following the species (or infraspecies) name can then distinguish between the two names or concepts. If databases do include authors of species names, then these should definitely be included in fields separate from the species' names themselves.

The concatenation of data where author names and dates are kept separate is usually not a major issue except in plants with autonyms (see below). Mixed databases of plants and animals, however may cause some problems where authorities are treated slightly differently. It should not present too many difficulties if the author fields are set up in the database to cater for these but rules for extraction may need to be different for the different Kingdoms.

Dalcin (2004) treats the authority as a label element under his nomenclatural data domain.

i. Data entry

With animal names the author name (usually in full) is always followed by a year; with plants, the author name or abbreviation is given alone.

Animals:
Emydura signata Ahl, 1932
Macrotis lagotis (Reid, 1937)
(the bracket indicates that Reid ascribed the species to a different genus)
Plants:
Melaleuca nervosa (Lindley) Cheel
synonym: *Callistemon nervosus* Lindley
(Lindley originally described it as a *Callistemon*; Cheel later transferred it to the genus *Melaleuca*).

With plants, occasionally the terms "ex" or "in" may be found in author names. The author in front of the "ex" - the pre-'ex' author is one who supplied the name but did not fulfil the requirements for valid publication or who published the name before the nomenclatural starting date for the group concerned. A post-'in' author is one in whose work a description or diagnosis supplied by another author is published. For a further explanation of pre-'ex' and post-'in' authors and their use see Arts 46.2 and 46.3 of the International Code of Botanical Nomenclature (2000). If author names are used within databases they should be in separate fields to the name (see discussion on atomisation, above) and it is recommended that neither the pre-'ex' nor the post-'in' authors be cited.

Green (1985) ascribed the new combination *Tersonia cyathiflora* to "(Fenzl) A.S. George"; since Green nowhere mentioned that George had contributed in any way, the combining author must be cited as "A.S.George ex J.W.Green" or preferably as just "J.W.Green".

Tersonia cyathiflora (Fenzl) J.W.Green

In W.T.Aiton's 2nd edition of *Hortus Kewensis* (1813), many of the descriptions are signed Robert Brown, and thus it can be assumed that Brown described the species. The author of the names is often cited as "R.Br. in Ait." It is recommended, however that the author be cited as just "R.Br."

Acacia acuicularis R.Br.

With plants – for the type subspecies or variety, etc. where the infraspecific name is the same as the species name (autonym), the author of the species name is used and follows the specific epithet. This format regularly causes confusion for reconstruction of names in specimen databases that include author names, as it is an exception to other rules.

Leucochrysum albicans (A.Cunn.) Paul G.Wilson subsp. *albicans*

For plants, abbreviation of authors' names follows an internationally agreed standard (Brummitt and Powell 1992), and this publication may be used to set up a checklist, or used for data entry and/or validation checking.

A.Cunn. = Allan Cunningham
L. = Linnaeus
L.f. = Linnaeus filius (son of-)

Sometimes, a space is given between Initial and Surname, others not. It is a matter of preference.

ii. Error checking

Author names as used in Brummit and Powell (1992) could be used to check authors in botanical database. Harvard University also has prepared a downloadable file of botanical authors and made this available on-line[2]. This should prove to be a very valuable file for checking authors' names and dates. Some names databases also include author names (e.g. IPNI 1999, Froese and Bisby 2002). Again Soundex-like techniques as mentioned above could be used to look for similarities between two names. It is the combination of species name and author that is the deciding factor, however, and these are not always easy to check.

If authors are used, then all published names in the database should have an author. In these cases, a Missing Data Values check should be carried out.

g. Collectors' names

Collectors' names are generally not standardised in collection databases, although standardisation of plant collector's names are being attempted for plant names in the *species*Link project in Brazil (Koch, 2003), and at Kew Gardens by Peter Sutton.

Extensive lists of collector's names have been published for some areas, but mainly for botanical collectors (see Steenis -Kruseman 1950, Hepper and Neate 1971, Dorr 1997, Index Herbariorum 1954-1988). There are also a number of on-line resources available:

- Harvard University have recently prepared a downloadable file of botanical collectors and collector teams and made these available on-line. http://www.huh.harvard.edu/databases/cms/download.html
- Index Collectorum – from the University of Göttingen http://www.sysbot.uni-goettingen.de/index_coll/default.htm
- Directory of Insect Collectors of Southern Africa (Entomological Society of Southern Africa http://www.up.ac.za/academic/entomological-society/collectr/collectr.html
- Index bio-bibliographicus notorum hominum Nonveilleriana (The Croatian Entomological Society) http://www.agr.hr/hed/hrv/bibl/osobe/comentsEN.htm

There are also a number of hard copy publications, and there are sure to be many more such indexes available in the various disciplines of zoology.

i. Data entry

It is recommended that names be included in primary species databases in a standard format. The HISPID Standard (Conn 2000) recommends the following:

> *Primary collector's family name (surname) followed by comma and space (,) then initials (all in uppercase and each separated by fullstops). All initials and first letter of the collector's family name in uppercase. For example, Chambers, P.F.*

It is recommended that secondary collectors be placed in a second field. If this is not the case, then it is recommended that they be cited with a comma and space used to separate the multiple collectors. For example:

> *Tan, F., Jeffreys, R.S.*

Where there is a chance of confusion, other given names should be spelt out. For example, to distinguish between Wilson, Paul G. and Wilson, Peter G. (a space is placed after the given name; no punctuation, except as separator between two names, as described above).

[2] http://www.huh.harvard.edu/databases/cms/download.html

Titles should be omitted.

If the family name (surname) consists of a preposition and a substantive, as in many European names (e.g. C.G.G.J. van Steenis), then the preposition is in lower case and the substantive has an initial capital letter. For example:

Steenis, C.G.G.J. van

Other names of similar form include de la Salle, d'Entrecasteaux, van Royen etc. It should be noted, however, that many of these names have been anglicised, particularly in America, such that both parts of the family name are treated as substantive. In such cases, these names can be transferred as follows:

De Nardi, J.C.

The prefixed O', Mac', Mc' and M' (e.g. MacDougal, McKenzie, O'Donnell) should all be treated as part of the substantive and hence transferred as part of the family name. For example:

McKenzie, V.

Hyphenated given names should be transferred as all uppercase, with the first and last initial separated by a hyphen (without spaces), and only the last terminated by a full stop. For example:

Quirico, A-L.

Peng, C-I.

If the collector of the record is unknown, then the term "Anonymous" should be used.

Interpreted information should be enclosed in square brackets, e.g.

Anonymous [? Mueller, F.]

ii. Error checking

As mentioned above, without a standard list of collectors, it is not easy to carry out error checks on collectors' names. This is particularly so in databases that do not follow a standard practice (such as putting surname first as mentioned above). If the database has standardised, then it is quite easy to sort all collector's names in the database and look for slight variations (for example a collector that uses one initial sometimes, and two at others). Extreme care should be taken not to introduce new errors to the database by altering a collector's name without absolute certainty that the change is correct. The initials example, above, is one case where a change could easily introduce new error. Errors that may be correctable are misspellings of surname, for example.

One way to develop a list of collectors is to create a list of unique values from the database in the same way as authority tables are developed for taxon names.

Fields associated with the *Collector-Name* such as *Date-of-Collection* may also be used for error checking. Historians have carried out a considerable amount of work recently on developing itineraries of explorers and collectors, historic scientific expeditions, ship itineraries, etc. Often these are not carried out by scientists, but by historians, and our science can benefit greatly from this work (see resources listed above, along with collections in the libraries (publications, journals, etc.) of many of the world's older museums and herbaria, recent work at the University of California, San Diego, the Scripps Institution of Oceanography, and the National Science Digital Library on capturing and documenting data from cruises as part of the SIO Digital Library Project). Links between those databases and primary species databases can lead to an improvement of both as inconsistencies and errors are detected.

Spatial Data

Spatial location is one of the most crucial aspects in being able to determine the fitness for use of many species-occurrence records. Spatially related biogeographic studies comprise one of the largest uses for these data – studies such as species distributional modelling, biogeographic studies, environmental planning and management, bio-regionalisation studies, reserve selection and conservation planning, and environmental decision support. For a detailed study, see the associated paper on *Uses of Primary Species-Occurrence Data* (Chapman 2005b).

We often think of primary species data as being ***point*** records of plant or animal occurrences but this is only part of the story. Seldom are collecting locations recorded accurately or precisely enough to be regarded as true points. The accuracy associated with the collection means that the point actually represents an area or a footprint. For example, a location from textual information that says "10 km north of Campinas", then there is an accuracy associated with the distance of "10 km" (perhaps ± 500 m), an accuracy associated with the direction "north" (i.e. north is somewhere between say NW and NE), and there is an accuracy associated with "Campinas" (is it the city boundary – a polygon – the city centre, etc.). For a more detailed discussion, see Wieczorek 2001a, Wieczorek *et al.,* 2004. In addition, many observational and survey records are recorded from an area (a ***polygon***) such as bird observations over a 2 ha area, or within a National Park, or from a regular grid (a ***grid***) such as observations from all 10-minute grid squares across Australia (Blakers *et al.* 1984), or from a 10 m X 10 m survey grid, or from along a transect (a ***line***) such as a transect survey or records along a road or river (although probably better treated as a polygon derived from buffering the road or river, depending on the scale). See further discussion under *Visualisation of Error* below.

As mentioned previously, a number of programs do exist that can aid in checking and testing for errors in geocodes attached to primary species records. Other tools are available to assist in the original assignment of coordinates to the data from the location information (such as distance and direction from a named location).

The testing of errors in already assigned georeferences involves

- Checking against other information internal to the record itself, for example, State or named district.
- Checking against an external reference using a database – e.g. is the record consistent with the collecting localities of the collector;
- Checking against an external reference using a GIS, including "point-in-polygon" tests – that the record falls on land rather than at sea, for example;
- Checking for outliers in geographic space; or
- Checking for outliers in environmental space.

Data Entry and Georeferencing

As stressed throughout these documents, error prevention is preferable to error detection, and the georeferencing or geocoding of records is one of the greatest sources of error in the databasing of species-occurrence data. Many new tools are now being developed to assist with the process of adding coordinates (especially latitude and longitude) to primary species data. This is not an easy process, however, especially as much of the legacy data (early collections in museums and herbaria collected over the past 300 or 400 years) carry little geographic information other than a general description of the location where they were collected (Chapman and Milne 1998). These collections were often made before modern settlements were built and named, and before roads were built. Many were collected from horseback or by boat, days from the last settlement and reference points were often difficult to determine. Many of the reference points no longer occur on modern maps and, in many cases where they do occur, they are ambiguous. Where geocodes are given, they are often not very accurate (Chapman 1999) and have generally been added at a later date (*retrospective georeferencing* – Blum 2001) by those other than the collector (Chapman 1992).

i. Definitions:

Before proceeding, there are a number of terms whose use in this document need definition. Some of these terms are used differently elsewhere, and different disciplines use different terms to define similar processes.

Geocode: As used in this paper, a geocode is the code (usually an x, y coordinate) that records the spatial positioning of a record according to a standard reference system. Here it is used to record a position on the surface of the earth. For species-occurrence data, the geocode is given according to one of several standard geographic reference systems with Universal Transverse Mercator, and latitude and longitude being two of the more common, and may be recorded in one of a number of ways (meters; decimal degrees; degrees, minutes, seconds; degrees, decimal minutes, etc.). Definitions of the term geocode are broad and wide-ranging. In many GIS applications it refers to an address and Zip Code, in marketing terms it refers to a demographic characterisation of a neighbourhood, and in some cases (Clarke 2002) it refers only to the location in "*computer readable form*". Also sometimes called a ***georeference*** or ***coordinate.***

Georeferencing: In this paper georeferencing is used to describe the process of assigning geographic coordinates to a record that links it to a geographic location on earth. Also sometimes called ***geocoding***.

ii. Database design

The design of databases for primary species-occurrence data should ensure that there are fields to properly cater for information that is often wrongly placed in the locality field – data such as habitat and habit information and geographic notes. An example of a distribution field with mixed information (from Fishbase[3] for *Perca fluviatilis*) is:

> "*Throughout Europe and Siberia to Kolyma River, but not in Spain, Italy or Greece; widely introduced. Several countries report adverse ecological impact after introduction*".

Such mixed fields are very difficult to treat in an automated way using parsing algorithms and are not consistent with the philosophy and design of relational databases where the information can be stored in Memo fields.

There are several additional fields that can be added to a species-occurrence database to assist in data cleaning and that can lead to an improvement in documenting data quality. Such fields include:

- ***Spatial accuracy*** – a field that records (preferable in meters, but sometimes in coded form) the accuracy with which a record's location has been determined.
- ***Named Place, Distance and Direction*** – some databases include "Nearest Named Place", "Distance" and "Direction" in separate fields as well as a plain text locality field. The inclusion of such fields can aid in geocode determination as well as in error checking.
- ***Geocode method*** – a field (or fields) that records how the geocode was determined – may include (Chapman 2005a)
  - use of differential GPS;
  - handheld GPS corrupted by Selective Availability (i.e. a recording prior to 1 May 2000);

[3] http://www.fishbase.org/

  - A map reference at 1:100 000 and obtained by triangulation using readily identifiable features;
  - A map reference using dead reckoning;
  - A map reference obtained remotely (eg. in a helicopter);
  - Obtained automatically using geo-referencing software using point-radius method;
  - Obtained from database using previously georeferenced locality.

- ***Geocode type*** – records the type of locality description that was used to determine the geocode.

  In a paper on the point-radius method of georeferencing locality descriptions, Wieczorek and others (2004) provide a table of nine types of locality descriptions found in natural history collections. The first three of these they recommend should not be georeferenced, but an annotation be given as to why it was not georeferenced. Some databases use a centroid with a huge accuracy figure (e.g. 100,000,000 meters). This has the drawback of users extracting the data only using the geocode and not the associated accuracy field and ending up with what looks like a point without its associated huge radius. The Wieczorek method overcomes this drawback by not providing such a misleading geocode. The nine categories listed by Wieczorek *et al.* (2004) are (with modified examples):

  1. ***Dubious*** (e.g. 'Isla Boca Brava?')
  2. ***Cannot be located*** (e.g. 'Mexico', 'locality not recorded')
  3. ***Demonstrably inaccurate*** (e.g. contains contradictory statements)
  4. ***Coordinates*** (e.g. with latitude or longitude, UTM coordinates)
  5. ***Named place*** (e.g. 'Alice Springs')
  6. ***Offset*** (e.g. '5 km outside Calgary')
  7. ***Offset along a path*** (e.g. '24 km N of Toowoomba along Darling Downs Hwy')
  8. ***Offset in orthogonal directions*** (e.g. '6 km N and 4 km W of Welna')
  9. ***Offset at a heading*** (e.g. 50 km NE of 'Mombasa')

Each of these would require a different method of calculation of the accuracy as discussed in the paper (Wieczorek *et al.* 2004

iii. Georeferencing Guidelines

Two excellent guidelines have been developed to assist data managers with georeferencing. The Georeferencing Guidelines developed by John Wieczorek at the Museum of Vertebrate Zoology in Berkeley (Wieczorek 2001) and the MaPSTeDI (Mountains and Plains Spatio-Temporal Database Informatics Initiative) guidelines (University of Colorado 2003) are two of the most comprehensive. I understand that there are also guidelines developed by Conabio in Mexico (CONABIO 2005), which are being translated into English, and thus will soon be available in both Spanish and English.

iv. Edit controls

Edit controls involve business rules that determine the permitted values for a particular field. One of the most frequent errors in spatial databases is the accidental omission of the '–'(minus) sign in records from the southern or eastern hemispheres. If the database is a database of all southern hemisphere records (a database of Australian records for example), then it should be automatic that all records are given a "negative" latitude. Databases of mixed records are, of course, more difficult to deal with, but the country and state fields could be used to check against the latitude or longitude.

Not all databases are set up correctly initially, and this can allow errors that should never occur. For example, latitudes greater than 90º or less than –90º and longitudes greater than 180º or less than –180º. If these are permitted by the database, then the database needs to be modified, otherwise checks need to be run on a regular basis to ensure that errors like these do not occur and are corrected.

v. Using existing databased records to determine geocodes

Information already included in the database can be used to assign georeferences to new records being added. A simple report procedure can be incorporated that allows for a search to ascertain if a specimen from the same locality has already been databased and assigned a geocode.

For example, you are about to database a collection that has the location information “10 km NW of Campinas” but no georeferencing information. You can search the database for “Campinas” and look through the collections already databased to see if a geocode has already been assigned to another collection from “10 km NW of Campinas”. This process can be made a lot simpler if the database structure includes fields for “Nearest Named Place”, “Distance” and “Direction” or similar, in addition to the traditional free text locality description.

This methodology has the drawback that if the first geocode had been assigned with an error, then that error will be perpetuated throughout the database. It does, however, allow for a global correction if such an error is found in any one of the collections so databased. If such a method is used to determine the geocode it should be so documented in the *Geocode method* field (see above).

With linked databases, such as the Australian Virtual Herbarium (CHAH 2002), *species*Link (CRIA 2002), or the GBIF Portal (GBIF 2004), on-line procedures could be set up to allow for a collaborative geocoding history to be developed and used in a similar way. Such collaboration may be carried out through the use of Web Services (Beaman *et al.* 2004, Hobern and Saarenmaa 2005). Of course, one drawback of this is that there is a certain amount of loss of control within your database, and an error in another database can be inadvertently copied through to your own database. Where this is done then the source-id should be attached to the record so that later updates and corrections can be incorporated. Good feed back mechanisms would need to be developed between institutions to ensure that, firstly errors were not perpetuated inadvertently, and secondly that information on errors that are detected are fed back to the originating database as well as other dependent databases.

Many plant collections are distributed as ‘duplicates’ to other collection institutions. Traditionally this has been done prior to georeferencing, and one can often find exactly the same collection in a number of different institutions, all with different georeferencing information. To circumvent these discrepancies, geocodes need to be added before distribution, or a collaborative arrangement entered into between institutions. As explained earlier, it costs a lot in both time and money to add geocodes, it is an extremely wasteful exercise if several institutions individually spend time and resources georeferencing the same collections. The waste is further compounded if different geocodes are given to the same collection in those separate institutions.

vi. Automated geocode assignment

Automated georeferencing tools are based on determining a latitude and longitude from the textual locality information using a distance and direction from a known location. Ideally, databases include at least a “Nearest Named Place”, “Distance” and “Direction”, or better still, “Named Place 1”, “Dist 1”, “Dir. 1”, “Named Place 2”, “Dist 2”, “Dir 2”. Thus “5 km E of Smithtown, 20 km NNW of Jonestown” would be appropriately passed into the six fields cited above.

As most databases are not so structured, attempts are being made to develop automated parsing software to parse free-text locality descriptions into basic “Nearest Named Place”, “Distance” and “Direction” fields, and then using these fields, in association with appropriate Gazetteers to determine the georeference (see *BioGeomancer* below). At the same time as the geocode is determined in this way, the Geocode Accuracy should be recorded in an extra field and where possible, the results checked by experts against the original to avoid unanticipated errors. In any case, such parsing should not in any way tamper with the original “Locality” data (field), but be additional information added. It can thus always be used to check the accuracy of the parsing exercise.

Drawbacks of this methodology include possible errors in the Gazetteers (most publicly available gazetteers have a considerable number of errors (see for example, figure 15), Nearest Named Place locations may refer to quite a large area (see comments below on assigning accuracy), many location fields are not as straight forward as those cited above, often historic place names are used, and many distances on collection labels are “by road” distances rather than direct, although this is seldom stated on the label itself. Accuracy fields need to take into consideration these issues as well as the error inherent in vector distances – does “South West” mean between “South” and “West” or between SSW and WSW. As this distance from the source increases, the inherent error in these will also rapidly increase (see discussion in Wieczorek *at al.* 2004). The use of this method in conjunction with a simple GIS would provide the opportunity for the operator to see the record on a map and to then “grab and drag” the point to a more appropriate place – for example to the nearest road.

vii. Geocoding software

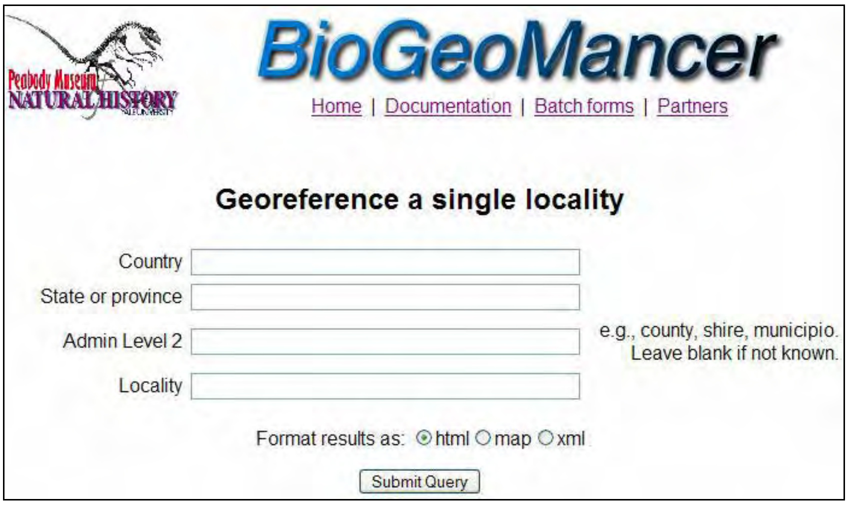
A number of on-line and stand-alone tools have been developed to assist users with geo-referencing their collections. Three are mentioned here – two ‘on-line’ and two ‘stand-alone’.

BioGeoMancer

BioGeoMancer is an automated georeferencing system for natural history collections (Wieczorek and Beaman 2002). In its present state, BioGeoMancer is a prototype system, and the comments below do not consider planned enhancements that are sure to improve its useability. BioGeoMancer can parse English language place name descriptions and provide a set of latitude and longitude coordinates associated with that description. The parsing of free-text, English language locality data provides an output of nearest named place, distance and direction, in the format (Wieczorek 2001b):

- 2.4 km WNW of Pandemonium
- Springfield, 22 miles E
- Springfield, 0.5 mi. E of Pandemonium

Like a number of other programs (e.g. Diva-GIS, eGaz) it takes the parsed information and in conjunction with an appropriate gazetteer, calculates a set of latitude and longitude coordinates. BioGeoMancer has the advantage over other geocoding programs in that it provides the parsing of the text. It is the first such geo-parsing program available to the public and researchers over the internet.



BioGeoMancer

Home | Documentation | Batch forms | Partners

Georeference a single locality

Country

State or province

Admin Level 2 — e.g., county, shire, municipio. Leave blank if not known.

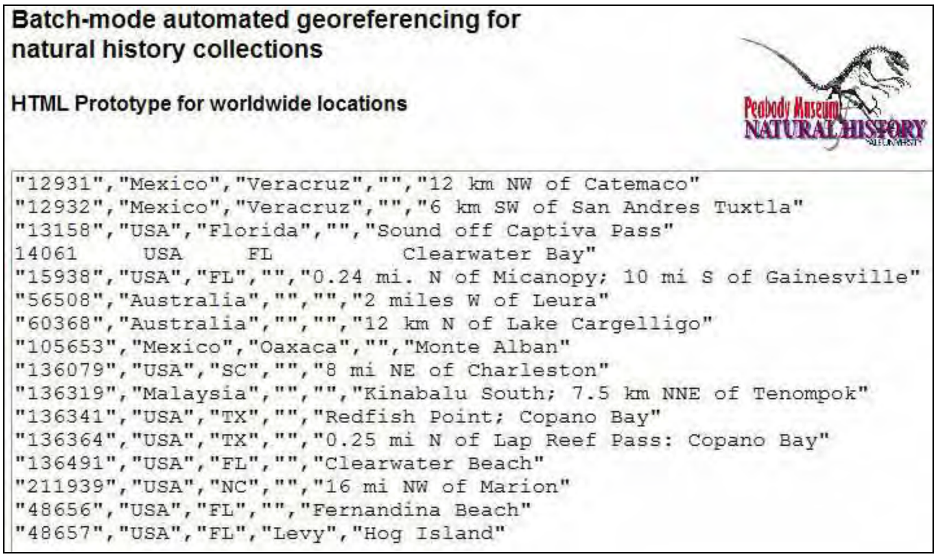
Locality

Format results as: ● html ○ map ○ xml

Submit Query

Fig. 2. *Single locality BioGeoMancer query form* *http://biogeomancer.org/* *(Peabody Museum n.dat.).*

The BioGeoMancer program exists in two forms. The first is a single specimen Web query form (figure 2) that allows the user to type in a locality and get a georeference returned. The second form, a batch process, accepts data through either an HTTP/CGI interface in a comma-delimited version (figure 3) or in a SOAP/XML version and provides a return file with georeferenced records either in delimited, html, table (figure 4), or xml format. This project has recently received a considerable boost in funding and expanded to become a worldwide collaboration attempting to develop new and improved georeferencing tools.

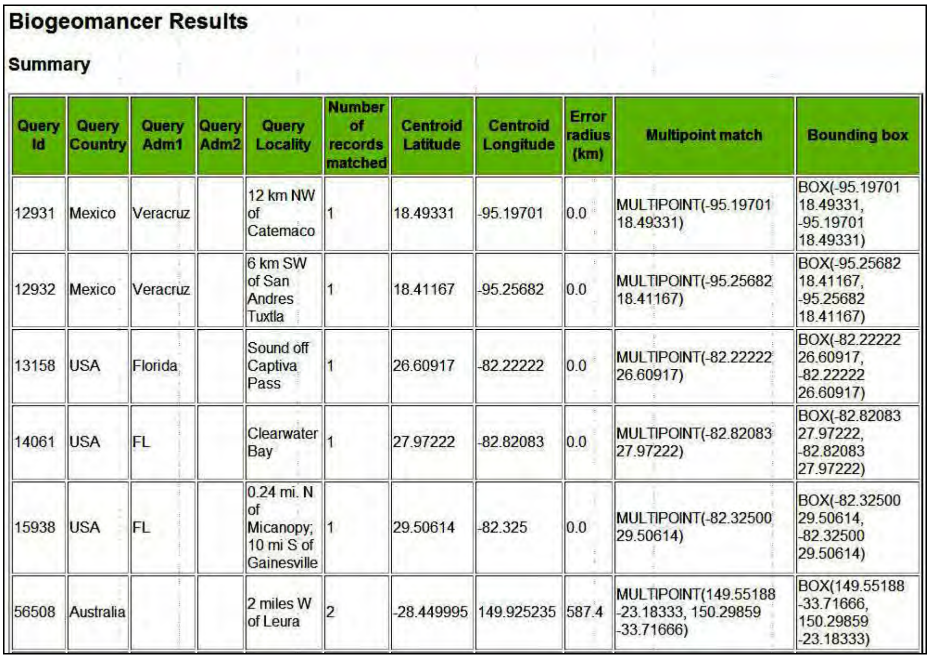


Batch-mode automated georeferencing for natural history collections

HTML Prototype for worldwide locations

```
"12931","Mexico","Veracruz","","12 km NW of Catemaco"
"12932","Mexico","Veracruz","","6 km SW of San Andres Tuxtla"
"13158","USA","Florida","","Sound off Captiva Pass"
14061     USA     FL          Clearwater Bay"
"15938","USA","FL","","0.24 mi. N of Micanopy; 10 mi S of Gainesville"
"56508","Australia","","","2 miles W of Leura"
"60368","Australia","","","12 km N of Lake Cargelligo"
"105653","Mexico","Oaxaca","","Monte Alban"
"136079","USA","SC","","8 mi NE of Charleston"
"136319","Malaysia","","","Kinabalu South; 7.5 km NNE of Tenompok"
"136341","USA","TX","","Redfish Point; Copano Bay"
"136364","USA","TX","","0.25 mi N of Lap Reef Pass: Copano Bay"
"136491","USA","FL","","Clearwater Beach"
"211939","USA","NC","","16 mi NW of Marion"
"48656","USA","FL","","Fernandina Beach"
"48657","USA","FL","Levy","Hog Island"
```

Fig. 3. *Input format for the BioGeoMancer web-based Batch-mode automated georeferencing tool for natural history collections* *http://georef.peabody.yale.edu/yu/bgm-forms/batch-int02.html* *(Peabody Museum n.dat.).*



Biogeomancer Results

Summary

Query Id	Query Country	Query Adm1	Query Adm2	Query Locality	Number of records matched	Centroid Latitude	Centroid Longitude	Error radius (km)	Multipoint match	Bounding box
12931	Mexico	Veracruz		12 km NW of Catemaco	1	18.49331	-95.19701	0.0	MULTIPOINT(-95.19701 18.49331)	BOX(-95.19701 18.49331, -95.19701 18.49331)
12932	Mexico	Veracruz		6 km SW of San Andres Tuxtla	1	18.41167	-95.25682	0.0	MULTIPOINT(-95.25682 18.41167)	BOX(-95.25682 18.41167, -95.25682 18.41167)
13158	USA	Florida		Sound off Captiva Pass	1	26.60917	-82.22222	0.0	MULTIPOINT(-82.22222 26.60917)	BOX(-82.22222 26.60917, -82.22222 26.60917)
14061	USA	FL		Clearwater Bay	1	27.97222	-82.82083	0.0	MULTIPOINT(-82.82083 27.97222)	BOX(-82.82083 27.97222, -82.82083 27.97222)
15938	USA	FL		0.24 mi. N of Micanopy; 10 mi S of Gainesville	1	29.50614	-82.325	0.0	MULTIPOINT(-82.32500 29.50614)	BOX(-82.32500 29.50614, -82.32500 29.50614)
56508	Australia			2 miles W of Leura	2	-28.449995	149.925235	587.4	MULTIPOINT(149.55188 -23.18333, 150.29859 -33.71666)	BOX(149.55188 -33.71666, 150.29859 -23.18333)

Fig. 4. *Sample partial output in tabular form from the BioGeoMancer web-based Batch-mode automated georeferencing tool for natural history collections (Peabody Museum n.dat.).*

Where more than one option is possible, then all are reported under that ID. Where no options are obvious, then the record is not returned. The *Bounding Box* column provides the calculated accuracy. The system works well for a lot of data, but does have difficulty with text that is not easily parsed into the above named place, distance and direction. Other noted issues in the current version include (future enhancements are planned that will reduce these):

- It is restricted to English-language descriptions.
- Accuracy is reported only as a bounding box in the present version, and this could be improved. Already, a related program developed by John Wieczorek (2001b) – the Georeferencing Calculator - can supply this information http://manisnet.org/manis/gc.html and this is likely to be linked to BioGeoMancer at a later date. Already work has begun on a method of assigning accuracy automatically through what has been termed the "point-radius method" for georeferencing and calculating associated uncertainty (Wieczorek *et al*. 2004)
- Two named localities (e.g. "10 km W of Toowoomba toward Dalby") produces a null result.

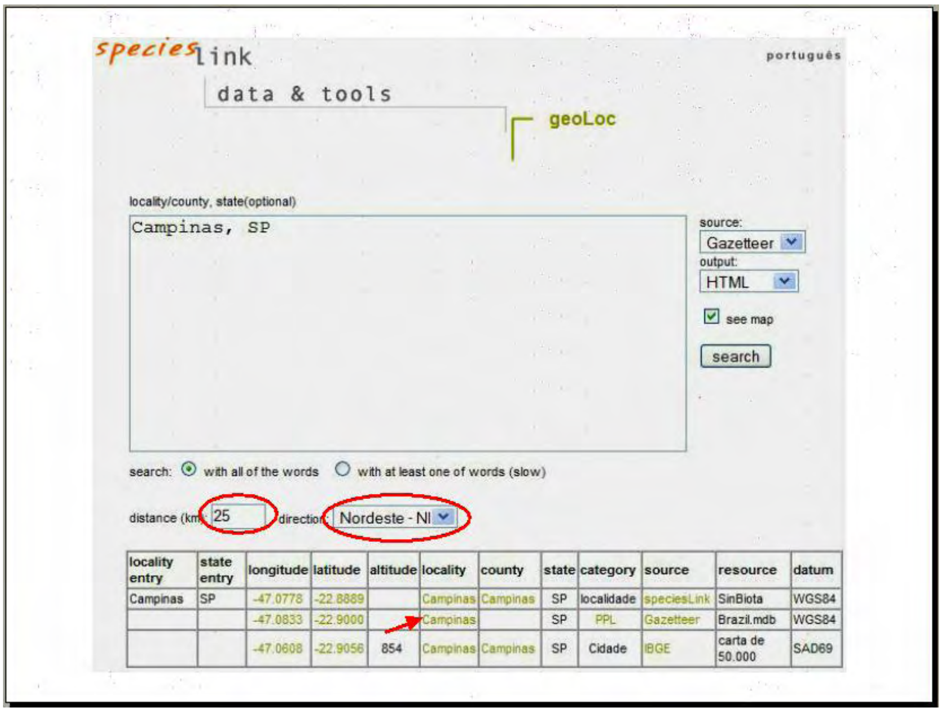
Another parsing program, RapidMap Geocoder (NMNH 1993) was developed in 1993 by the US National Museum of Natural History and the Bernice P. Bishop Museum in Hawaii, for use only with Hawaiian localities. However it was not considered successful and was discontinued. Some useful information on the parsing methodologies used, however, is available on the internet at: http://users.ca.astound.net/specht/rm/tr_place.htm.

GeoLoc-CRIA

GeoLoc is a simple web-based program for finding localities in Brazil, a known distance and direction from a gazetted locality. It has been developed at CRIA (Marino *et al.* in prep.). GeoLoc works in a similar way to the eGaz program (see below) and can be found at http://splink.cria.org.br/tools/ (CRIA 2004a). The prototype includes a number of gazetteers and provides the user with the potential to select which gazetteer if more than one is available for an area, and also provides a calculated error value.

An example can be seen in figure 5, where the latitude and longitude of a locality 25 km NE of Campinas in São Paulo, Brazil is sought. Firstly one finds the locality for Campinas using one of a number of Gazetteers, or the *species*Link records (records obtained through distributed searching of a range of databases mainly in the State of São Paulo). Then by adding “25 km” and “NE” (circled) and clicking on the relevant ‘Campinas’ (arrow), the results will appear on an associated map (figure 6). The geocode is given as -46.9244, -22.7455 with an error of 9.754 km (circled). This information (latitude, longitude and error) are already stored in the Microsoft paste buffer and can be pasted into any Microsoft Windows compatible file such as Word, Excel, and Access. The map also shows the location of “Campinas” from the three sources – the one in red being the one chosen, along with the point “25 km NE of Campinas”. The map can be zoomed and panned, and various environmental layers turned on or off.

The program can also link to an EXCEL spreadsheet of localities and produce an html table of results for further searching, or an EXCEL spreadsheet. The main drawback of the program is that it is only available for use with Brazilian locations. The algorithms are currently being incorporated into the wider Biogeomancer project.



locality/county, state(optional)

Campinas, SP

source: Gazetteer

output: HTML

☑ see map

search

search: ● with all of the words ○ with at least one of words (slow)

distance (km): 25 direction: Nordeste - NE

locality entry	state entry	longitude	latitude	altitude	locality	county	state	category	source	resource	datum
Campinas	SP	-47.0778	-22.8889		Campinas	Campinas	SP	localidade	speciesLink	SinBiota	WGS84
		-47.0833	-22.9000		Campinas		SP	PPL	Gazetteer	Brazil.mdb	WGS84
		-47.0608	-22.9056	854	Campinas	Campinas	SP	Cidade	IBGE	carta de 50.000	SAD69

Fig. 5. *Using CRIA’s ‘geoLoc’ program to find the geocode for a locality 25 km NE of Campinas, SP.*

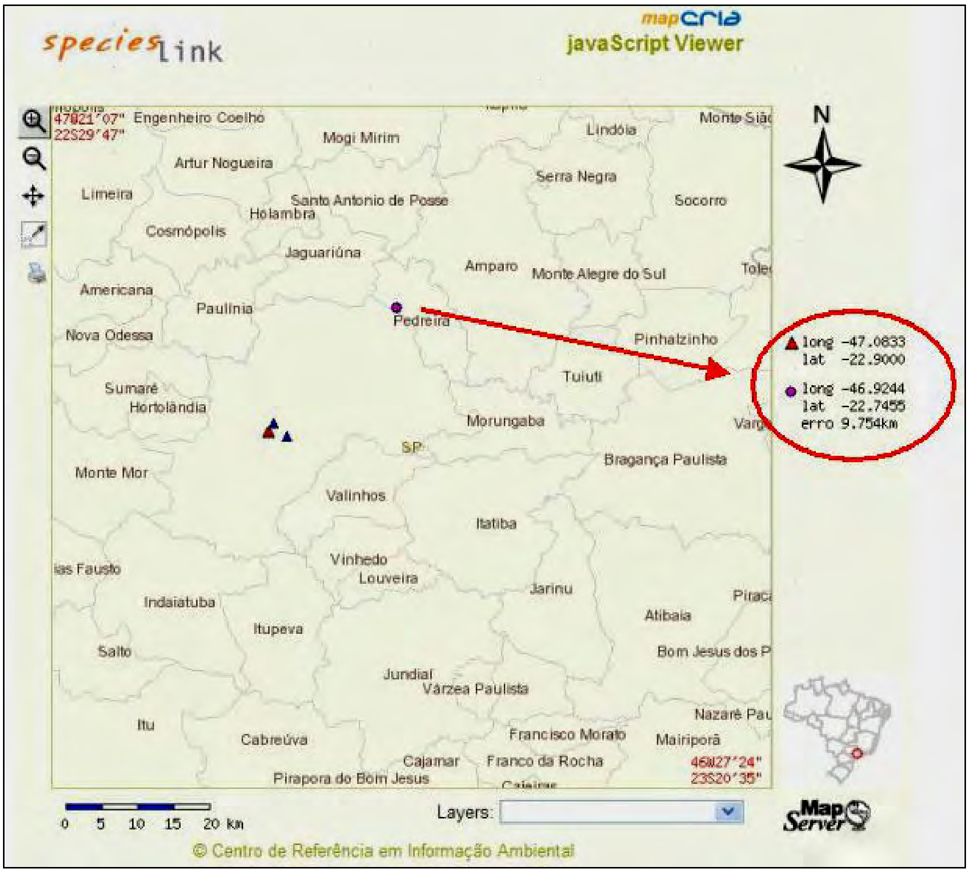


Fig. 6. *Results of the above selection showing the location of "Campinas" (from the various sources) and the point 25 km NE of Campinas, with associated geocode information and error (circled)*

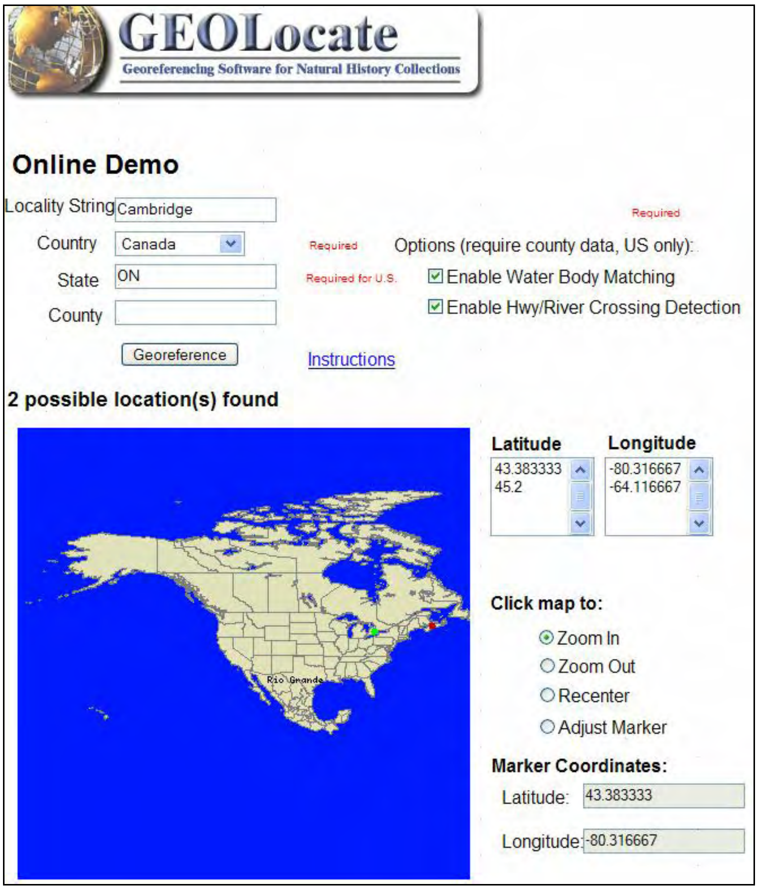
GEOLocate

GEOLocate (Rios and Bart *n.dat*.) is a georeferencing program developed by Tulane University's Museum of Natural History and is designed to facilitate the task of assigning geographic coordinates to the locality data associated with natural history collections. The primary goals of GEOLocate are to:

- develop an algorithm to convert textual natural history data into latitude and longitude for North America;
- provide an interface for visualisation and further adjustment of generated coordinates;
- provide a simple solution for users to import and georeference their data;
- provide an auto-updating feature.

The algorithm first standardises the locality string into common terms and parses out distances, direction and key geographic identifiers such as the named place. This information is then used in conjunction with gazetteers (including placenames, river miles, landuse and road/river crossing data) to determine the geographic coordinates. The program also allows the user to "snap" localities to the nearest water-body.

The program is available from the University of Tulane, and an on-line demonstration is available at: http://www.museum.tulane.edu/geolocate/demo.aspx (figure 7).



GEOLocate
Georeferencing Software for Natural History Collections

Online Demo

Locality String: Cambridge
Country: Canada
State: ON
County:

Options (require county data, US only):
☑ Enable Water Body Matching
☑ Enable Hwy/River Crossing Detection

2 possible location(s) found

Latitude	Longitude
43.383333	-80.316667
45.2	-64.116667

Click map to:
● Zoom In
○ Zoom Out
○ Recenter
○ Adjust Marker

Marker Coordinates:
Latitude: 43.383333
Longitude: -80.316667

Fig. 7. *An example of the GEOLocate interface using the on-line demo version to identify the geographic coordinates for Cambridge, Ontario.*

The program only works for North America (Mexico, USA and Canada), but the developers are currently working on extending it to include the entire world. Other developments will include DiGIR compatibility, multi-lingual support, and advanced validation techniques (N.Rios pers. com. 2004).

eGaz

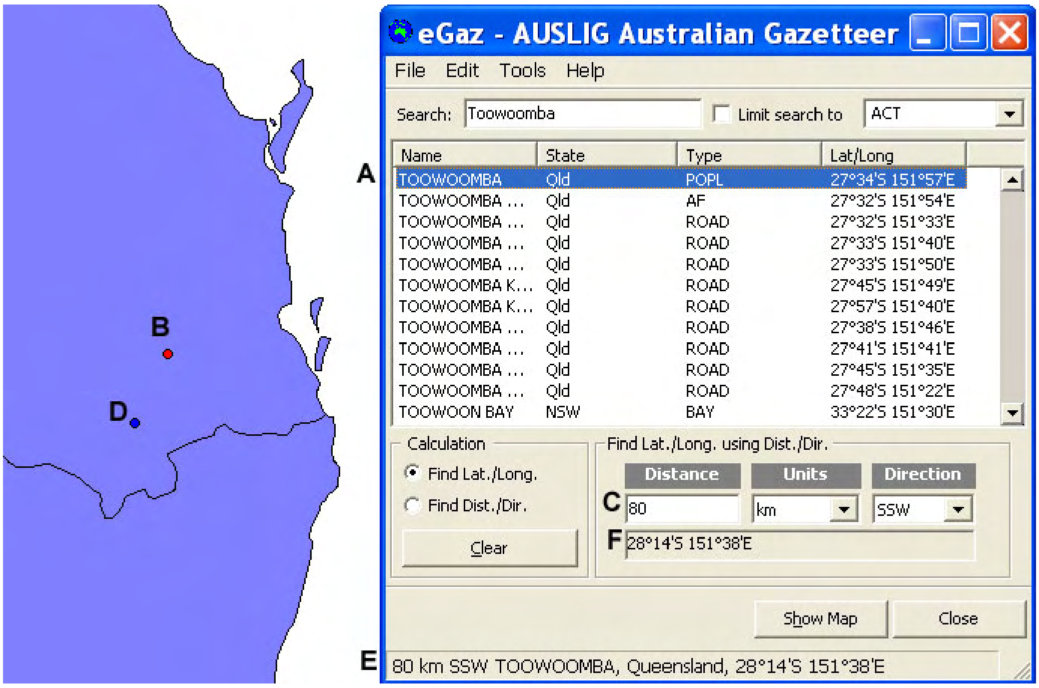
eGaz (Shattuck 1997) is a program developed at the CSIRO's Australian National Insect Collection to assist museums and herbaria to identify and add geocodes to their specimen records. With the development of the data entry and specimen management software, BioLink (Shattuck and Fitzsimmons 2000), it was incorporated into that software package. eGaz is available as part of the Biolink package (see below).

eGaz eliminates the need for paper based maps and rulers to determine the latitude and longitude for cities, towns, mountains, lakes and other named places. eGaz can also calculate latitude and longitude for sites a known distance and direction from a named place. The program allows for the easy inclusion of Gazetteers from any region, and Gazeteers for much of the world are available for download from the CSIRO site (http://www.biolink.csiro.au/gazfiles.html).

eGaz is a Microsoft Windows based product that provides two windows, a Gazetteer window and a Map window (figure 8). It allows the user with a location in the form of a "Named Place", "Distance" and "Direction" to obtain a geocode and transfer that to a file.

The example shown in figure 8 is of obtaining the latitude and longitude of a position "80 km SW of Toowoomba", Queensland, Australia. The first step is to load the appropriate Gazetteer and select “Toowoomba” from it (**A**). There are a number of options, but I have selected the Toowoomba City (labelled POPL for Populated Place). The location of Toowoomba appears on the map in red (**B**). The distance “80” is typed into the Distance field and the pull down menus used to select “km” and “SSW” (**C**). The selected location appears on the map as a blue dot (**D**). The location, along with the latitude and longitude also appears on the bottom of the Gazetteer window (**E**). By right clicking on this area and selecting “Copy” that information can be copied and pasted into any Microsoft Windows compatible file (Word, Excel, Access). The Latitude and Longitude (to 1 arc-minute resolution) also appears (**F**), and this can similarly be copied to a file. Alternatively, by going to the Edit menu and select “Copy Lat/Long” the geocode can be copied to an accuracy of one arc-second.

One can also go to the map itself and zoom in to the point. Other layers such as a road network (in ESRI Shape file format) can be loaded to allow more accurate positioning of the point – i.e. perhaps move it to the nearest road if collecting was done from a vehicle. The selection tool can then be used to click on the point to obtain the geocode to one arc-second resolution. Again right clicking with the mouse, or using Edit/Copy Lat/Long, allows the information to be copied to an appropriate file.


B
D
A
C
F
E
eGaz - AUSLIG Australian Gazetteer
File
Edit
Tools
Help
Calculation
Find Lat./Long. using Dist./Dir.
Clear
Show Map
Close

Search: Toowoomba □ Limit search to ACT

Name	State	Type	Lat/Long
TOOWOOMBA	Qld	POPL	27°34'S 151°57'E
TOOWOOMBA ...	Qld	AF	27°32'S 151°54'E
TOOWOOMBA ...	Qld	ROAD	27°32'S 151°33'E
TOOWOOMBA ...	Qld	ROAD	27°33'S 151°40'E
TOOWOOMBA ...	Qld	ROAD	27°33'S 151°50'E
TOOWOOMBA K...	Qld	ROAD	27°45'S 151°49'E
TOOWOOMBA K...	Qld	ROAD	27°57'S 151°40'E
TOOWOOMBA ...	Qld	ROAD	27°38'S 151°46'E
TOOWOOMBA ...	Qld	ROAD	27°41'S 151°41'E
TOOWOOMBA ...	Qld	ROAD	27°45'S 151°35'E
TOOWOOMBA ...	Qld	ROAD	27°48'S 151°22'E
TOOWOON BAY	NSW	BAY	33°22'S 151°30'E

● Find Lat./Long.
○ Find Dist./Dir.

Distance: 80
Units: km
Direction: SSW

28°14'S 151°38'E

80 km SSW TOOWOOMBA, Queensland, 28°14'S 151°38'E

Fig. 8. *Sample output from eGaz, showing the determination of latitude and longitude for a position 80 km SSW of Toowoomba, Queensland, Australia.* ***A.*** *Information on Toowoomba from Gazetteer.* ***B.*** *Mapped location of Toowoomba.* ***C.*** *Input showing 80 km SSW of highlighted location.* ***D.*** *Mapped location 80 km SSW of Toowoomba.* ***E.*** *Details on location.* ***F.*** *Latitude and Longitude of new location.*

Diva-GIS

Diva-GIS is a free GIS program developed for use in museums and herbaria. It includes an algorithm that assists in assigning coordinates to specimen data where this is lacking. Some pre-processing is necessary to organise the data into a format acceptable to the program, but a number of databases are already beginning to structure their data in this way. The input file requires the textual location data to be parsed into a number of specialised fields. These are “Named Place1”, "Distance 1", "Direction 1" and "Named Place2", "Distance 2", "Direction 2". For example the locality record:

> "*growing at a local place called Ulta, 25.2 km E of Chilla*"

would be parsed to:

Named place 1:	Ulta
Distance 1:	
Direction 1:	
Named Place 2:	Chilla
Distance 2:	25.2km
Direction 2:	E

and

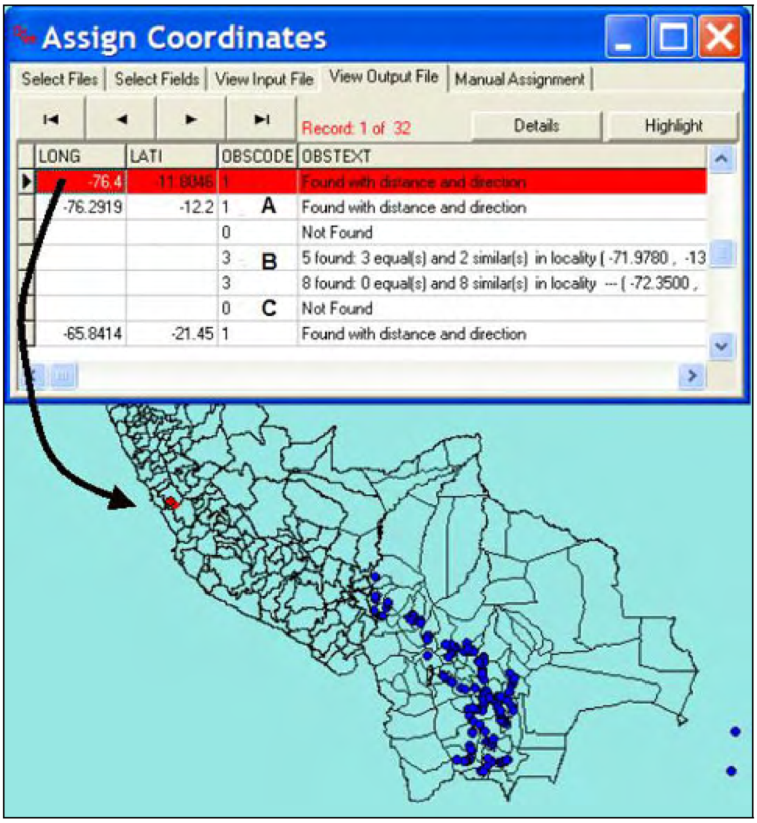
> "14 km ESE of Sucre on road to Zudanez"

would parse to:

Named place 1:	Sucre
Distance 1:	14 km
Direction 1:	ESE
Named Place 2:	Zudanez
Distance 2:	
Direction 2:	

Just one set of "Named Place", "Distance" and "Direction", however, will be able to provide the geocoding for many records, and this is all the information most institutions will have. The authors of the Diva-GIS (Hijmans *et al.* 2005) recommend rounding the distance down to whole numbers to account for inaccuracies in the data, and to cater for cases where 25 km North of a place, really means 25 km North by road and not in a direct line. I would recommend to the contrary, and would record the most accurate figure given, and place an accuracy figure in an "Accuracy" field in meters.

Once an input file has been selected, an output file named, and the appropriate field names selected from a pull-down list, the algorithm is run and produces an output file (figure 9). The algorithm uses an appropriate Gazetteer to assign coordinates.

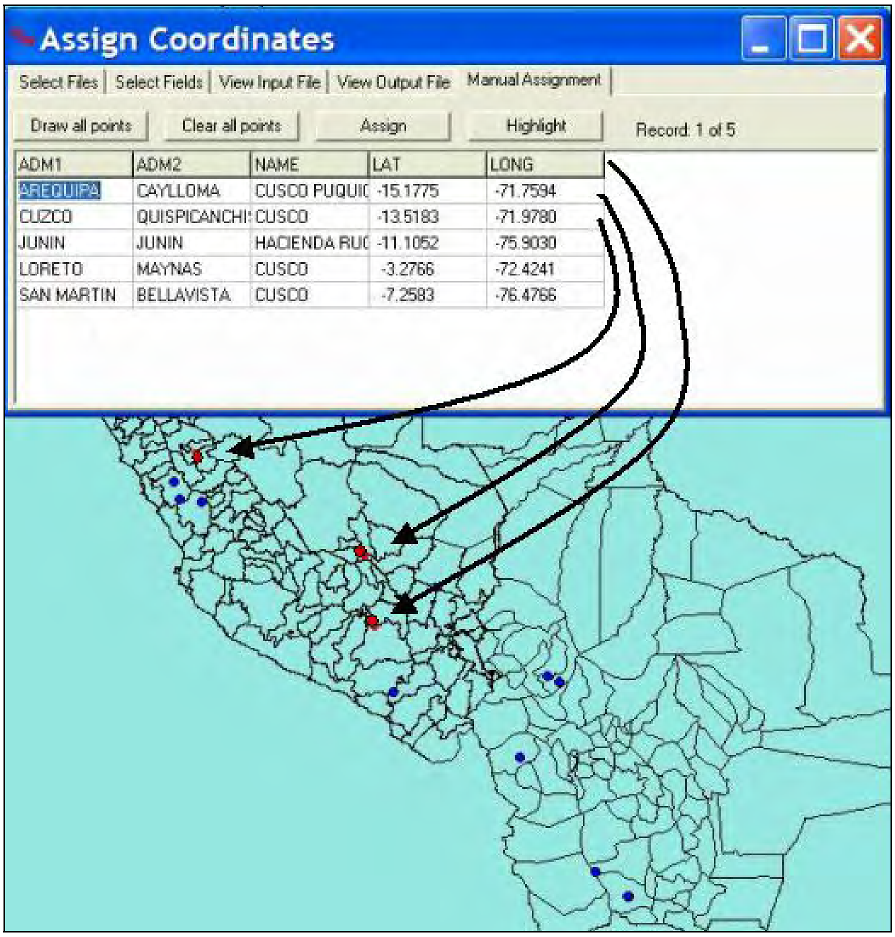


LONG	LATI	OBSCODE	OBSTEXT
-76.4	11.8046	1	Found with distance and direction
-76.2919	-12.2	1	Found with distance and direction
		0	Not Found
		3	5 found: 3 equal(s) and 2 similar(s) in locality (-71.9780 , -13
		3	8 found: 0 equal(s) and 8 similar(s) in locality --- (-72.3500 ,
		0	Not Found
-65.8414	-21.45	1	Found with distance and direction

Fig. 9. *Results from Diva-GIS showing point records with geocodes automatically assigned.* ***A.*** *Unambiguous geocodes found by the program and assigned.* ***B.*** *Ambiguous geocodes identified.* ***C.*** *Appropriate geocodes not found.*

As shown in the example (figure 9), the program has found unambiguous matches in the Gazetteer(s) for a number or records using the "Named Place" field in the input file and assigned those records an appropriately calculated geocode (**A**). Once the output file has been loaded and a shape file created, each of these records can be highlighted to produce a flashing point on the map. In a number of other cases, the program has found several possible matches in the Gazetteer(s) for the "Named Place" and reported on that appropriately (**B**). In yet other cases (**C**) the program has been unable to find a match in the Gazetteer.

In the case of records where a number of possible matches were found, one can go to the next stage by double clicking on one of the (**B**) records and producing another output file (figure 10).



ADM1	ADM2	NAME	LAT	LONG
AREQUIPA	CAYLLOMA	CUSCO PUQUI	-15.1775	-71.7594
CUZCO	QUISPICANCHI	CUSCO	-13.5183	-71.9780
JUNIN	JUNIN	HACIENDA RU	-11.1052	-75.9030
LORETO	MAYNAS	CUSCO	-3.2766	-72.4241
SAN MARTIN	BELLAVISTA	CUSCO	-7.2583	-76.4766

Fig. 10. *Results from Diva-GIS showing alternate geocodes for a record where use of the Gazetteer has produced a number of credible alternatives.*

In the case of the record shown in figure 11, the program has identified five possible alternative locations from the Gazetteer(s) and presented these alternatives on the GIS for the user to choose. When one is chosen, it is just a matter of clicking on the “Assign” button for that to be assigned to the output file. Alternatively, one can decide on another location altogether and use the “Manual Assignment” to add a geocode or modify one of the assigned ones.

Geocode checking and validation

There are four main methods that can be used for checking and validating geocodes on specimen records once databased. These are the use of databases for checking internal inconsistencies, the use of geographic information systems, the use of environmental space to check for outliers and the use of statistics to check for outliers in geographic or environmental space.

i. Using Databases

a. Internal checks

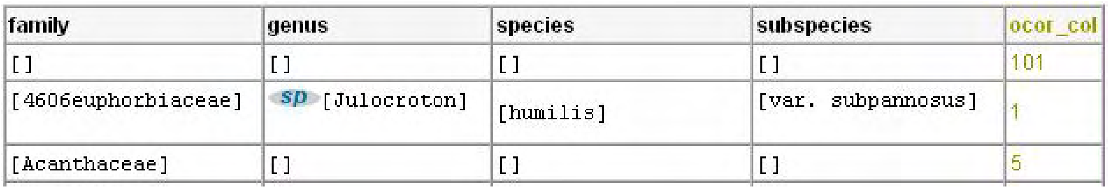
Most species and species-related databases include a certain amount of redundant information. For example, the State in which the collection was made as well as a field for textual location information. Some databases also include a "nearest named place" and this may also duplicate information within the locality field. Checks can then be made to check that the cited town or nearest named place in one field, is located within the correct State or district, or even country as cited in another field.

Checking information in a database between similar records is also possible, for example, checking all localities against the supplied latitude and longitude. One may have a database with 5 collections from one location – "10 km N of Campinas, SP" for example. Do they all have the same latitude and longitude or are one or more significantly different to the others? See also discussion on *Ordinal Association Rules* below.

Data Cleaning (speciesLink)

CRIA's Data Cleaning module of the speciesLink Distributed Information System (CRIA 2002) includes a number of routines for identifying possible errors to help collection managers in processing their data. At the moment this is only in Portuguese, but an English version is proposed. One portion of this tool is to identify errors in names. Routines include:

- Listing of all names (family, genus, species, subspecies) along with the number of occurrences in the databases accessed. A brief look at one example (figure 11) shows a number of obvious problems. The first line shows that there are 101 occurrences in the database with records not identified at any level from family below. The second line shows one occurrence with a family name "4606euphorbiaceae", and line 3 shows 5 records in Acanthaceae identified to family only.

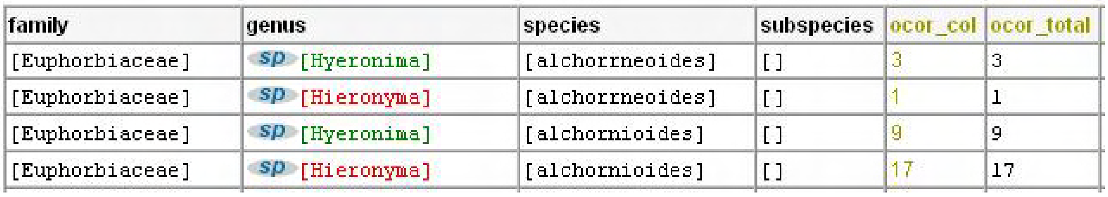


family	genus	species	subspecies	ocor_col
[]	[]	[]	[]	101
[4606euphorbiaceae]	sp [Julocroton]	[humilis]	[var. subpannosus]	1
[Acanthaceae]	[]	[]	[]	5

Fig. 11. *Extract from CRIA Data Cleaning module showing some possible errors.*

- Examining possible errors in generic names. This is where the family name is the same, the species name is the same, the generic names are similar (identified using soundex-like algorithms) but the spelling of the generic name is different. This output also shows the number of occurrences of each in the database being studies, and the total number of occurrences in all databases accessed through speciesLink. The example (figure 12) shows two different spellings of the genus *Hieronyma* (along with two spellings of *alchornioides*, but those are identified under a different routine) along with the number of occurrences of each. One can click on the "sp" and it takes you to a search of a range of databases both internal to the organisation as well and external, and includes such resources as the

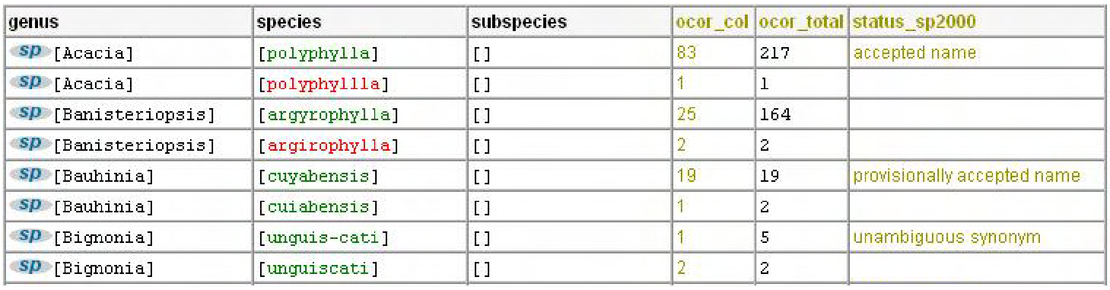
International Plant Name Index (IPNI), species 2000, etc. which can all help the user determine which may be the correct spelling.



family	genus	species	subspecies	ocor_col	ocor_total
[Euphorbiaceae]	sp [Hyeronima]	[alchorneoides]	[]	3	3
[Euphorbiaceae]	sp [Hieronyma]	[alchorneoides]	[]	1	1
[Euphorbiaceae]	sp [Hyeronima]	[alchornioides]	[]	9	9
[Euphorbiaceae]	sp [Hieronyma]	[alchornioides]	[]	17	17

Fig. 12. *Extract from CRIA Data Cleaning module showing some possible errors.*

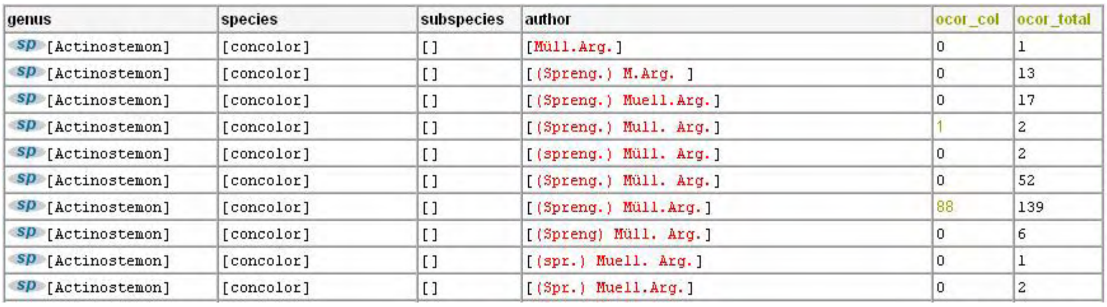
- Examining possible errors in species names or epithets. Like the generic names, this looks for names where the genus is the same, the soundex for the species epithet is the same, but there is a difference in the spelling of the species epithet. Again the output shows the total number of occurrences in the database being studied, the total occurrences in all databases accessed and the status of the name in species2000[4]. The example (figure 13) shows a number of species names with alternatives. The number of occurrences of each name along with the status in species2000 if available can give an indication of which of the spellings may be an error.



genus	species	subspecies	ocor_col	ocor_total	status_sp2000
sp [Acacia]	[polyphylla]	[]	83	217	accepted name
sp [Acacia]	[polyphyllla]	[]	1	1	
sp [Banisteriopsis]	[argyrophylla]	[]	25	164	
sp [Banisteriopsis]	[argirophylla]	[]	2	2	
sp [Bauhinia]	[cuyabensis]	[]	19	19	provisionally accepted name
sp [Bauhinia]	[cuiabensis]	[]	1	2	
sp [Bignonia]	[unguis-cati]	[]	1	5	unambiguous synonym
sp [Bignonia]	[unguiscati]	[]	2	2	

Fig. 13. *Extract from CRIA Data Cleaning module showing some possible errors.*

- Examining differences and possible errors in author names. Figure 14 shows the number of possibilities for just one species name. Again clicking on the "sp" a search of other databases can be carried out to help determine which may be the best alternative to use.



genus	species	subspecies	author	ocor_col	ocor_total
sp [Actinostemon]	[concolor]	[]	[Müll.Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spreng.) M.Arg.]	0	13
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Muell.Arg.]	0	17
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Mull. Arg.]	1	2
sp [Actinostemon]	[concolor]	[]	[(spreng.) Müll. Arg.]	0	2
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll. Arg.]	0	52
sp [Actinostemon]	[concolor]	[]	[(Spreng.) Müll.Arg.]	88	139
sp [Actinostemon]	[concolor]	[]	[(Spreng) Müll. Arg.]	0	6
sp [Actinostemon]	[concolor]	[]	[(spr.) Muell. Arg.]	0	1
sp [Actinostemon]	[concolor]	[]	[(Spr.) Muell.Arg.]	0	2

Fig. 14. *Extract from CRIA Data Cleaning module showing some possible errors.*

- Examining differences in family names and in subspecies names works in a similar manner.

Other routines are used to identify possible geographic errors in the datasets, and these are treated under *Spatial Data* below. CRIA is not a custodian of the data, and makes no changes to the data, but provides a service to data custodians, to help them identify possible errors in their databases. It is then up to the custodians to decide which are the correct formats and which records should be corrected and which not.

[4] http://www.species2000.org

b. External databases

By linking to external databases, errors in various aspects of the species-occurrence data can be identified. Such databases can include Digital Elevation Models, spatial topographic databases, gazetteers and collector's itineraries.

More sophisticated databases can be used to check the accuracy of the altitude fields by comparing the altitude cited with that of a databased Digital Elevation Model (DEM). It is important that the DEM used be at an appropriate scale, and due to the varying accuracy of most specimen data, can lead to false or misleading errors if not used critically. Such a technique has been used successfully in ERIN (Environmental Resources Information Network) in Australia for over 10 years (Chapman unpublished). The process uses batch processing using an ORACLE® database and can check (or assign) altitude records to over 3000 records a minute.

More recently, sophisticated spatial databases have been developed such as ESRI's Spatial Database Engine (ArcSDE®) (ESRI 2003) and PostGIS that allow for more complicated database searching using the geocodes themselves. This type of software, however, is very expensive, and very few museums or herbaria are likely to afford them or have the need for them and for that reason, these methods are not further outlined further in this paper.

Gazetteers exist for most of the world in one form or another, and frequently these are available as a downloadable database. They can be used to check appropriate fields within the specimen database for accuracy. Care needs to be exercised with the use of many of these databases as often they, themselves, contain errors (see for example figure 15), and it is important that the right gazetteer for the area, at an appropriate scale is used. Also, many named place names may be ambiguous (e.g. there are hundreds of "Sandy Creek"s in Australia) (Chapman and Busby 1994), or involve historic place names that do not occur in the modern gazetteer. There is also the issue of what a place name may actually mean (Wieczorek 2001a). One of the aspects of the new BioGeomancer project (see comments elsewhere) is the integration of Gazetteers with biological databases using Web Service technology. It is also hoped to improve gazetteers through public participation, and to especially begin including historic collection locations.

One method that is seldom used, but that has great potential, is cross checking against databases of collectors' localities. To date, very few such databases exist (but the Harvard database referenced above would be a good general starting point for botany[5]), and others are gradually being developed. Peterson *et al.* (2003) recently suggested a novel statistical method using the birds of Mexico as an example. They ordered the collections of a particular collector in temporal order and for each day (or group of days) impose a maximum radius of likely movement. Using a formula-based approach in EXCEL, they identified possible errors in specimens that fall outside the calculated range. Similar methods to this could be carried out in the database itself – see discussion under *Ordinal Association Rules*, below. Such a method will only work, however, if the databased collections from the collector are large enough to create such an itinerary.

ii. GIS Checks

Geographic Information Systems (GIS) are very powerful tools that have become much more user friendly in recent times. GISs range from expensive, high functionality systems to free, off-the-shelf products with more limited functionality. Many of the free GISs are powerful enough, however, to provide much of the functionality required by a herbarium or museum, and can be easily adapted to provide a range of data checking and data cleaning routines.

[5] http://www.huh.harvard.edu/databases/cms/download.html

	Points	Lines	Polygons
Points	▪ is a neighbour of ▪ is allocated to	▪ is near to ▪ lies on	▪ is a centroid of ▪ is within
Lines		▪ crosses ▪ joins	▪ intersects ▪ is a boundary of
Polygons			▪ is overlain by ▪ is adjacent to

Table 3: *Relationships between classes of objects (from Gatrell 1991)*

The GIS can also be used to check for logical consistency within the database. Redundancy in topological encoding can be used to detect flaws in data structure such as missing data and unlabelled polygons (Chrisman 1991). GIS allows the interrelation of spatial layers to detect errors and that, along with visualisation, is its major strength.

The use of a simple GIS to plot points (specimen records) against polygons (regions, States, Countries, soils) can aid in detecting mismatches in the data (either geographic or altitudinal). This is a common test used in GIS systems and is known as the "point-in-polygon" method – it is used in GIS to make sure marine buoys don't occur on land, that rivers don't occur outside their flood plains, etc. One of the most important tests a GIS can perform on primary species data is to check that records that are supposed to be on the land actually are on land, and those that are supposed to be in the ocean, are. It is obvious when one first loads a large data set into a GIS that many records are obviously in the wrong place just from this simple check. Checks for misplaced records using a GIS can range from simple visual inspection to more automated checking. Visual inspection can also be valuable in determining if records fall in the correct country, for example. If you have a database of records from Brazil, by using a GIS you can quickly identify records that are misplaced in such a way that they are outside of Brazil. For example, in figure 15, records from a publicly available Gazetteer of Brazilian place names have some obvious errors. Errors in specimen records can similarly be identified using this methodology.

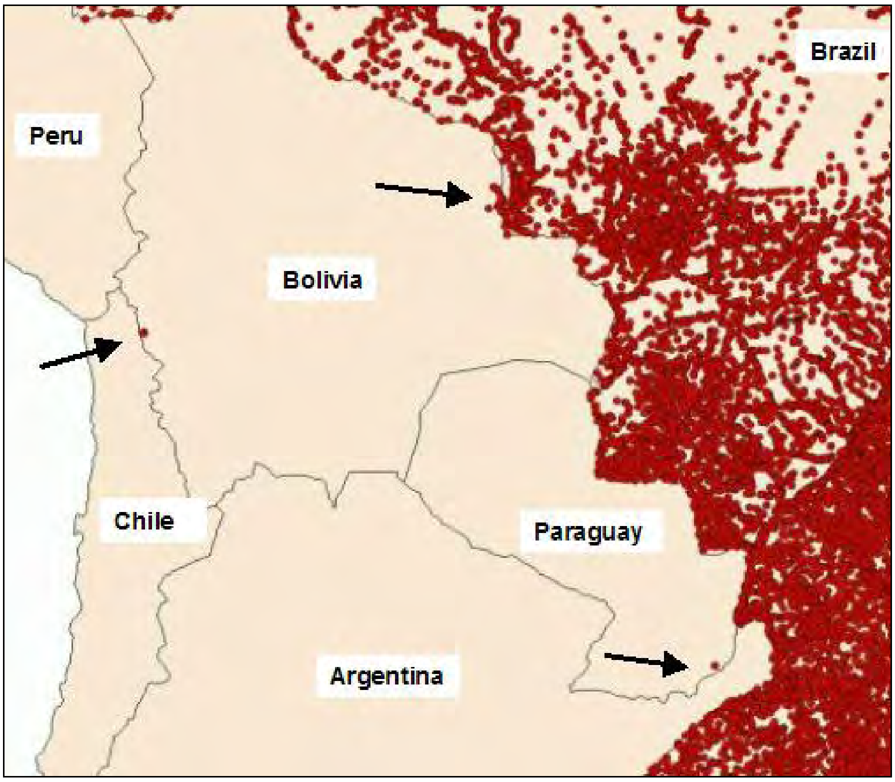


Fig. 15. *Records from a Gazetteer of Brazilian place names showing a number of errors (arrowed), with one obvious error sitting on the Chile-Bolivian border and another in southern Paraguay.*

A number of the tools, for example Diva-GIS (Hijmans *et al*. 2005) and the CRIA Data Cleaning tool (CRIA 2005) have routines that assist in identifying such errors.

The GIS can also be used to check that records fall outside a particular vegetation type, soil type or geology, etc. Some species are highly specific to certain geological types - limestone, sandstone, serpentenite (figure 16), for example. If you have the boundaries of these, any record that falls outside may be regarded as a possible outlier and flagged for further checking (Chapman *et al.* 2001). In figure 16, a species that only occurs on highly mineralised Serpentenite soils is mapped and two records (marked 'a' and 'b') show up as likely errors. On checking, record 'a' only has the locality 'Goomeri' – the nearest town to the Serpentenite outcrop, and has been geocoded with the latitude and longitude of the town. Record 'b' is quite near the outcrop and is likely misplaced due to the precision of the geocode given.

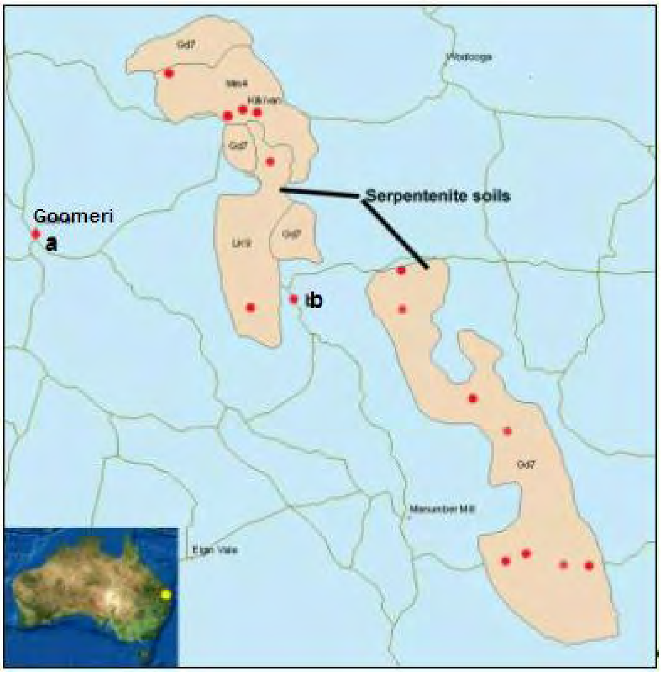


Fig. 16. *Records of a species (red) that is only found on highly mineralised Serpentenite soils. Records marked 'a' and 'b' have likely errors in geocoding.*

The identification of collectors' itineraries (Chapman 1988, Peterson *et al.* 2003), allows for checking for possible error if, for example, the date of collection doesn't fit the particular pattern of that collector. This could be particularly useful for collectors from the 18th and 19th centuries prior to collectors being able to cover vast distances within the one day using helicopters, planes or motor vehicles. In the example in figure 17, collections between 22 and 25 February and in the first half of March should be in the Pentland-Lolworth area (circled), if outside that, they are likely to include errors in the date of the collection, or in the geocode (Chapman 1988). Again, using a GIS to map both the itinerary and the species' records can be very valuable. Another example is the use of animated GIS in Nepal to trace the routes of collectors along rivers (Lampe and Reide 2002).

Other uses of a GIS include for example, buffering of likely locations – e.g. streams for fish and aquatic plants, the coast for littoral species, altitudinal ranges for alpine species or others known to have a distinct altitudinal range. In this way, anything outside the buffer may need to be checked. Care needs to be exercised, as with the fish, for example, it may mean that those records outside the buffer zone are not errors at all, but the species may be occurring in small streams too small for mapping. These tests can generally only flag suspect records, and then it is up to individual checks to determine what may be real errors in the record, and what may be true outliers.

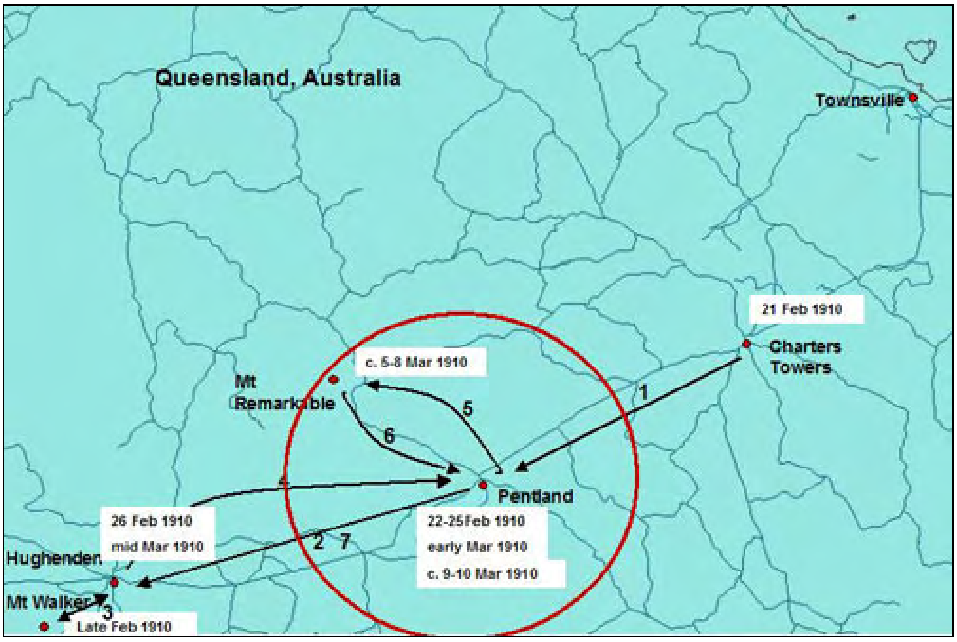


Fig. 17. *Collecting localities of Karl Domin in Queensland, Australia in 1910 (Chapman 1988). He travelled by train from Townsville to Hughenden, stopping at Charters Towers and Pentland. He then returned and spent about 10 days in the Pentland, Mount Remarkable, Lolworth area on horseback, before returning to Hughenden by train. Dates are only approximate.*

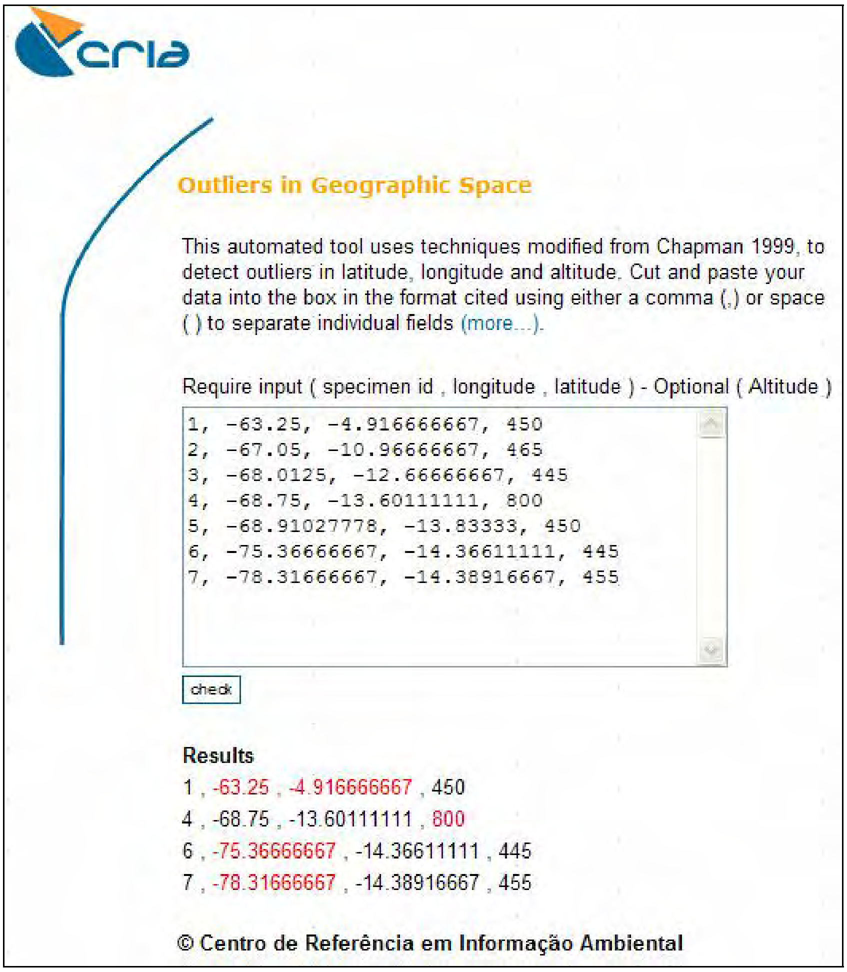
iii. Outliers in geographic and environmental space

There are a number of methods for detecting outliers in data and these are outlined below. Natural history data is very diverse and generally does not conform to standard statistical distributions, and thus, as suggested by Maletic and Marcus (2000), more than one method is often necessary to capture most of the outliers.

Geographic Outlier Detection

A program from CRIA in Brazil (spOutlier) allows a user to type or cut and paste records into a box on the internet, link to a file, or submit an XML file of specimen records and receive information on geographic outliers. Records are submitted in the form: “id, latitude, longitude, altitude” and the program returns information on likely errors, both in textual form and on a map interface (Marino *et al*. in prep). It also allows the user to identify their data set as either an on-shore (terrestrial) or off-shore (marine) and again the program will return a list of mismatches. This is a unique program, and one that will prove very useful to biologists. It is also possible for users to submit a document on-line and have it returned, annotated with information on possible errors. An on-line version can be seen at http://splink.cria.org.br/tools/ (CRIA 2004b).

In figure 14, the list of localities have returned four records with possible errors, 3 with possible errors in latitude, one with a possible error in longitude and one with a possible error in altitude. These points are then shown on the associated map with the records with possible errors identified in red.



Outliers in Geographic Space

This automated tool uses techniques modified from Chapman 1999, to detect outliers in latitude, longitude and altitude. Cut and paste your data into the box in the format cited using either a comma (,) or space ( ) to separate individual fields (more...).

Require input (specimen id , longitude , latitude) - Optional (Altitude)

```
1, -63.25, -4.916666667, 450
2, -67.05, -10.96666667, 465
3, -68.0125, -12.66666667, 445
4, -68.75, -13.60111111, 800
5, -68.91027778, -13.83333, 450
6, -75.36666667, -14.36611111, 445
7, -78.31666667, -14.38916667, 455
```

Results

1 , -63.25 , -4.916666667 , 450

4 , -68.75 , -13.60111111 , 800

6 , -75.36666667 , -14.36611111 , 445

7 , -78.31666667 , -14.38916667 , 455

© Centro de Referência em Informação Ambiental

Fig. 18. *Shows the prototype Outliers in Geographic Space system at CRIA identifying records 1, 4, 6 and 7 as having possible errors in geocoding.*

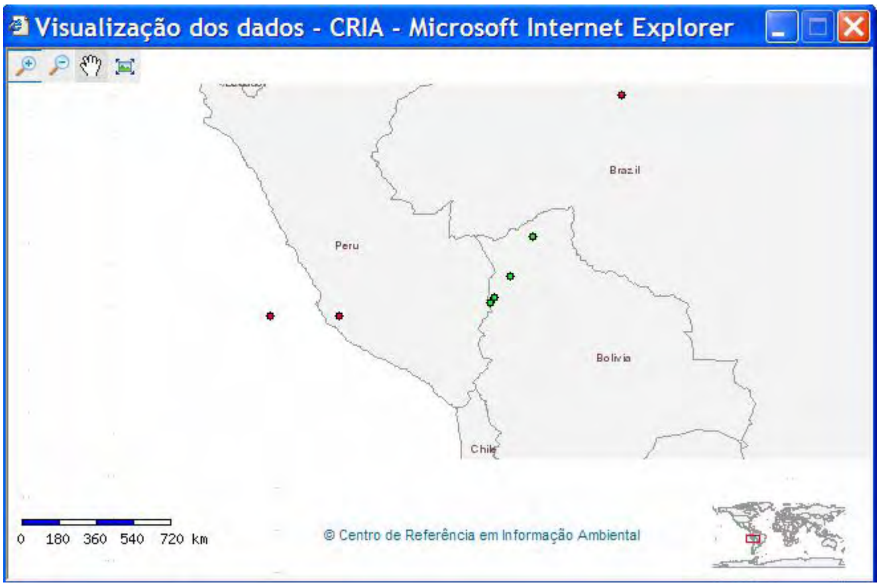

Visualização dos dados - CRIA - Microsoft Internet Explorer
Brazil
Peru
Bolivia
Chile
0 180 360 540 720 km
© Centro de Referência em Informação Ambiental

Fig. 19. *Map output associated showing identified suspect records (in red) from figure 14.*

Publicly available programs using this method:

- **spOutlier-CRIA** (CRIA 2004b, Marino *et al*. in prep).
- **Data Cleaning-CRIA** (CRIA 2005).

- **Diva-GIS** (Hijmans *et al.* 2005)

Cumulative Frequency Curves

Early versions of the program BIOCLIM (Nix 1986, Busby 1991) were used to detect possible outliers by excluding records that fall outside the 90 percentile range of any element of the climate profile for the taxon, or by using cumulative frequency curves (Busby 1991, Lindemeyer *et al.* 1991) where the percentile figure can be varied. Although these techniques are still in use and are easy to use (Houlder *et al.* 2000, Hijmans *et al.* 2005) they do not allow for taxa that may not include any genuine errors, or that include many errors. They are also suspect for very small sample sizes (Chapman and Busby 1994, Chapman 1999).

A recent modification of the Diva-GIS software (Hijmans *et al.* 2005) has lead to the inclusion of the Reverse Jackknifing methodology (Chapman 1999) discussed below, and this has been linked to the Cumulative Frequency Curve with records identified under that method highlighted on the Cumulative Frequency curve for each parameter.

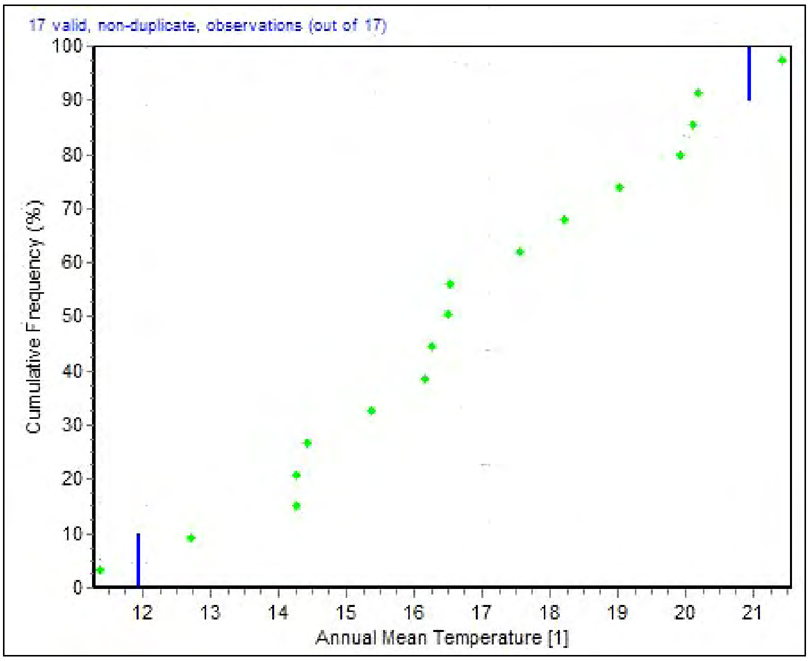


Fig. 20. *Cumulative frequency curve used to detect outliers in climate space using Annual Mean Temperature. The Blue lines represent the 97.5 percentile, the point on the bottom left (or even the two to the bottom left), may be regarded as a possible outlier worth checking for error in the geocode.*

Publicly available programs using this method:

- **Diva-GIS** (Hijmans *et al.* 2005)
- **ANUCLIM** (Houlder *et al.* 2000).

Principle Components Analysis

By using the scatter of points in a Principal Components Analysis of one climate layer against another one can identify possible outliers and thus possible errors in geocoding. It is a fairly powerful data validation method but unless the process is automated in some way to identify multiple outlier records the method can be quite tedious as one has to flick through however many combinations of climate components one is using.

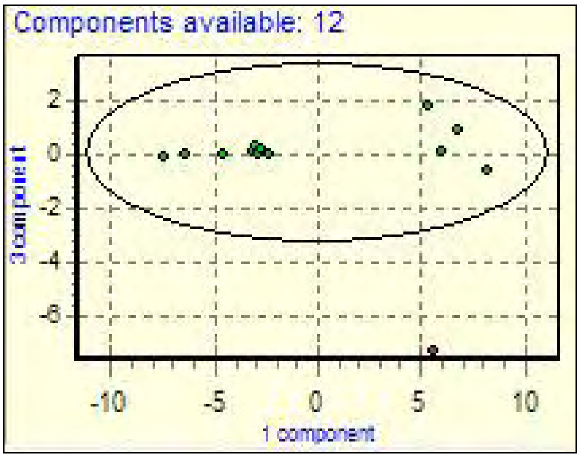


Fig. 21. *Principal Components Analysis, showing one point (in red) identified as an outlier and thus a possible error (from FloraMap, Jones and Gladkov 2001).*

Publicly available programs using this method:

- **FloraMap** (Jones and Gladkov 2001)
- **PATN** vers. 3.01 (Belbin 2004)

Cluster Analysis

The identification of outliers using clustering based on Euclidian or other distance measures can sometimes identify outliers that are not identified by methods at the field level (Johnson and Wichern 1998, Maletic and Marcus 2000). Cluster Analysis can be used to help identify multiple groups of like populations (using climate space or some other criteria), and can thus also be used to identify clusters that are isolated as either unicates or small groups separated by a significant distance from other clusters. Again, it is quite a valuable and seemingly robust methodology, but can depend very much on the cluster method used and can be computationally complex).

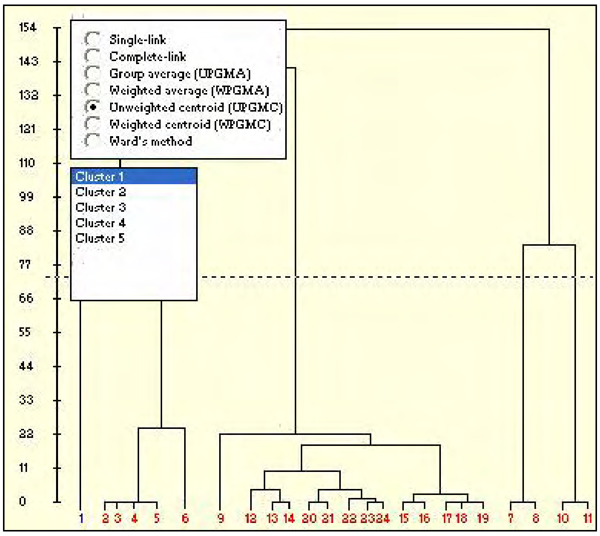
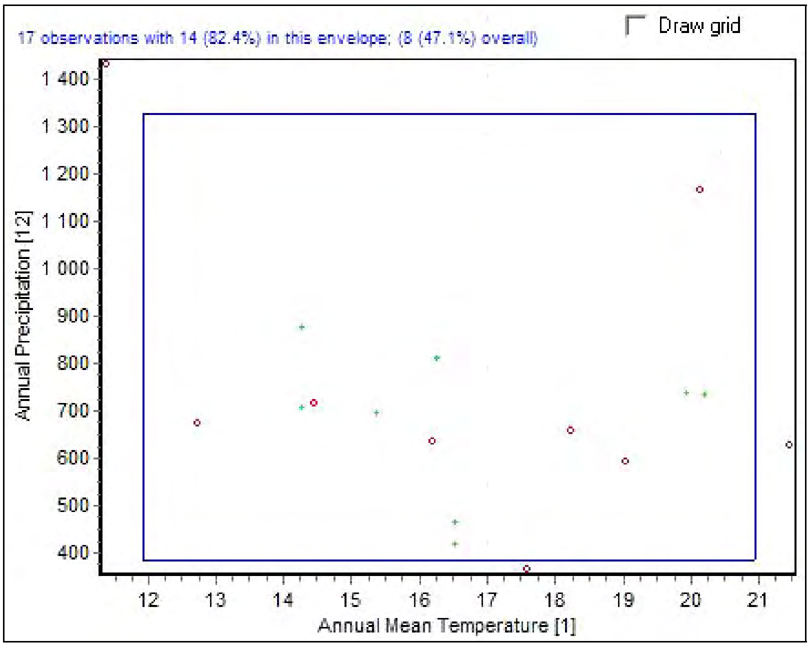

○ Single-link
○ Complete-link
○ Group average (UPGMA)
○ Weighted average (WPGMA)
● Unweighted centroid (UPGMC)
○ Weighted centroid (WPGMC)
○ Ward's method

Fig. 22. *Cluster Analysis showing a unicate cluster (#1 – in blue) which may be regarded as an outlier (from FloraMap, Jones and Gladkov 2001).*

Publicly available programs using this method:

- **FloraMap** (Jones and Gladkov 2001)
- **PATN** Vers. 3.01 (Belbin 2004)

Climatic Envelope


17 observations with 14 (82.4%) in this envelope; (8 (47.1%) overall)

□ Draw grid

Fig. 23. *Climatic envelope from BIOCLIM using a 97.5 percentile envelope for annual mean temperature and annual mean rainfall. Records marked in red are records that fall outside any one of the 64 possible envelopes.*

The Climatic Envelope method is an extension of the cumulative frequency curve methodology mentioned above, but groups each of the climate layers into a multi-dimensional box or envelope that can be examined two dimensions at a time, similar to the principal components analysis.

Outliers in any of the cumulative frequency curves that make up the totality of climate layers can be identified in this manner.

Publicly available programs using this method:

- **Diva-GIS** (Hijmans *et al.* 2005)

Reverse Jackknife

This technique uses a modified reverse jackknifing to extract outliers at either end of an array of points in any one of a number of climate profiles. In 1992, the method was developed at ERIN in Australia to automatically detect outliers in climate space (Chapman 1992, 1999, Chapman and Busby 1994) and thus identify suspect records amongst the thousands of species being modelled at the time. The method has proved extremely reliable in automatically identifying suspect records, with a high proportion (around 90%) of those identified as being suspect, proving to be true errors.

$$x < \bar{x}$$

if

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(\bar{x} - x_{(i)}\right)$$

else

$$y_{(i)} = \left(x_{(i+1)} - x_{(i)}\right)\left(x_{(i+1)} - \bar{x}\right)$$

then

$$C = \frac{y_{(i)}}{\sqrt{\frac{\sum_{i=1}^{n}\left(y_{(i)} - \bar{y}\right)^2}{n-1}}}$$

Fig. 24. *Formula for determining the Critical Value (C) in an outlier detection algorithm where C = Critical Value (from Chapman 1999). This formula has been used in Australia since 1992 for detecting outliers in environmental (climate) space. The formula has recently been modified (2005) by dividing the value of C by the range of 'x' and has been incorporated into Diva-GIS version 5.0 (Hijmans et al. 2005). This has improved its reliability for use with criteria with large values such as rainfall, elevation, etc.*

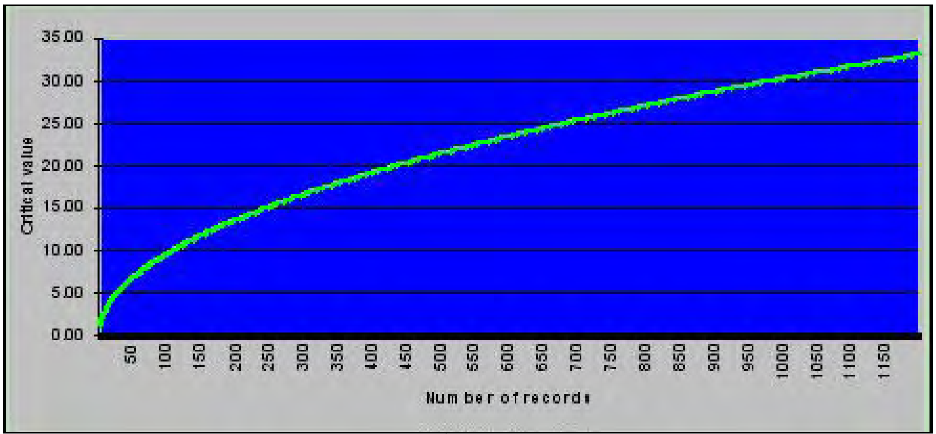
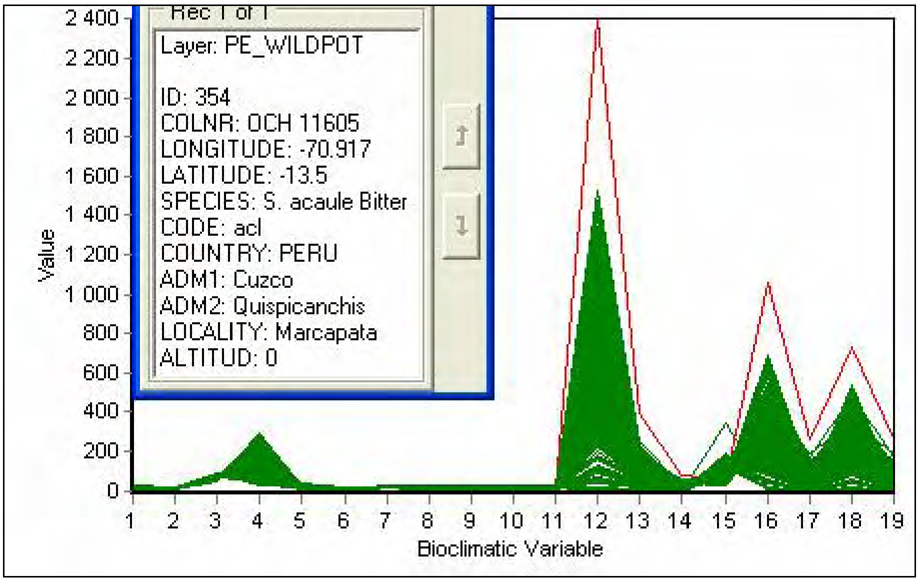


Fig. 25. *Threshold Value Curve (T=0.95(√n)+0.2 where 'n' is the number of records). Values above the curve are regarded as "suspect", values below the curve as "valid" (from Chapman 1999).*


Rec 1 of 1

Layer: PE_WILDPOT

ID: 354
COLNR: OCH 11605
LONGITUDE: -70.917
LATITUDE: -13.5
SPECIES: S. acaule Bitter
CODE: acl
COUNTRY: PERU
ADM1: Cuzco
ADM2: Quispicanchis
LOCALITY: Marcapata
ALTITUD: 0

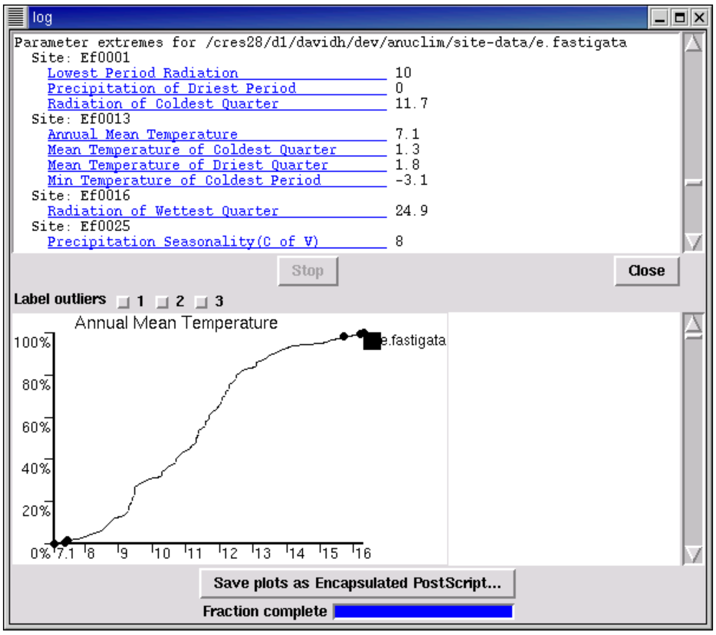
Fig. 26. *Outlier Detection algorithm in Diva-GIS using Reverse Jackknifing. The program has identified one possible outlier (using the selected option to show only records that were outliers in at least 6 (of 19 possible) criteria).*

Publicly available programs using this method:

- **Diva-GIS** (Hijmans *et al.* 2005)
- Also being programmed into the new BioGeomancer toolkit to be available mid 2006

Parameter Extremes

Parameter Extremes is a similar method to the Climatic Envelope method and identifies the record at the extremes of each Cumulative Frequency curve and bundles them into an output log file. In this way one can identify particular records that are extremes in more than one climate parameter.



```
Parameter extremes for /cres28/d1/davidh/dev/anuclim/site-data/e.fastigata
  Site: Ef0001
    Lowest Period Radiation                        10
    Precipitation of Driest Period                 0
    Radiation of Coldest Quarter                   11.7
  Site: Ef0013
    Annual Mean Temperature                        7.1
    Mean Temperature of Coldest Quarter            1.3
    Mean Temperature of Driest Quarter             1.8
    Min Temperature of Coldest Period              -3.1
  Site: Ef0016
    Radiation of Wettest Quarter                   24.9
  Site: Ef0025
    Precipitation Seasonality(C of V)              8
```

Fig 27. *Log file of Eucalyptus fastigata from ANUCLIM Version 5.1 (Houlder et al. 2000) showing the parameter extremes (top) and one associated species accumulation curve (bottom).*

Publicly available programs using this method:

- **ANUCLIM** (Houlder *et al.* 2000).

Other Methods

Many of the methodologies listed below are simple and are available in many standard statistical packages. Some do not seem to have been used for detecting errors in biological data, but examples with similar types of data indicate that they may be worth trying. A number of these and other methods are elaborated in Legendre and Legendre (1998). A number of other outlier detecting methods that may be worth trying, can be found in the publication by Barnett and Lewis (1994).

i. Standard Deviations from the Mean

Perhaps the most promising of these other methods would be to look at a varying number of standard deviations from the mean based on Chebyshev's theorem (Barnett and Lewis 1994). Maletic and Marcus (2000) tested a number of deviations from the mean using 5000 naval personnel records, with 78 fields of the same type (dates) and found that using 5 times the Standard Deviation generated the best results. Testing would need to be carried out on a number of collection datasets, and especially testing with much lower numbers of records than that used by Maletic and Marcus. Preliminary tests with low numbers by myself using elevation has so far not looked promising.

ii. Deviations from the Median

Another group of non-parametric statistical tests use relationships with the median rather than the mean. Two possible methods are the Mann-Whitney U test and the Kuskall-Wallis test which look at the alternate hypothesis that two (Mann-Whitney), or three or more (Kuskall-Wallis), populations differ only in respect to the median (Barnett and Lewis 1994, Lowry 2005). I have not seen examples of these applied to the detection of outliers in species-occurrence data, but they may be worth testing.

iii. Use of Modelled Distributions

Distribution models derived from species distribution modelling such as those produced using GARP (Stockwell and Peters 1999, Pereira 2002) or Lifemapper (University of Kansas 2003b), could be used to identify new records that fall outside the predicted distribution. This method, although easy to use, is limited by the quality of the predicted distribution. If all records for a species had not been used to develop the model, then there may be deficiencies in that model. Also, using just the outer boundaries of the distribution does not take into account the scattered nature of good models that identify suitable niches within the broad totality of the geographic distribution.

iv. Pattern Analysis

Pattern Analysis can be used to identify records that do not conform to existing patterns in the data. A variety of methods can be used for analysis of the patterns, including Association, Partitioning, Classification, Clustering, Ordination and use of Networks such as minimum spanning trees (Belbin 2004). Some of these methods have been discussed in more detail above. A pattern can generally be defined as a group of records that have similar characteristics (Maletic and Marcus 2000), but the choosing of the “right reference pattern” if such exists, can have an influence on the results (Weiher and Keddy 1999).

Publicly available programs using this method:

- **PATN** (Belbin 2004)

v. Ordinal Association Rules

Association rules attempt to find ordinal relationships that tend to hold over a large percentage of records (Marcus *et al.* 2001). They can be used for both categorical data and quantitative data. Simply put, they look for patterns such as if A<B most of the time, then if A>B in a record, then it is likely to be an error. With quantitative data, the rules can be used in conjunction with other statistical methods that use the mean, median, standard deviation and percentile ranges for outlier detection. With this method, the larger the number of records, the better the results, and in many cases could be used across whole databases rather than just within one species record. Uses could be such as, if species A occurs in Vegetation type B most of the time, then a record that has the information that it occurs in Vegetation type C may be an error. Or all records collected by a collector should not be within 15 years of the collector’s birth date, or greater than 100 years of their birth date, or later than their death date. Such rules could also be used in conjunction with a collector’s likely range (see above). For example, if a collection was before 1900, then two collections collected on the same day should not be greater than x kilometres apart.

Publicly available programs using this method:

- **PATN** (Belbin 2004).

Descriptive Data

Checking for errors in Descriptive Data is more difficult to cover here because of the quite diverse nature of what may be included in such databases. The structured nature of these databases, however, allow for more rule setting when the databases are set up.

i. Database design

The key to maintaining good data quality with descriptive databases is to follow good design procedures, and where possible design the databases following standards such as DELTA (Dallwitz *et al*. 1993) or the new SDD (Structure of Descriptive Data) standard (http://160.45.63.11/Projects/TDWG-SDD/) that is being developed by the Taxonomic Databases Working Group (TDWG).

ii. Edit controls

Because of the structured nature of descriptive databases, they lend themselves to the use of edit controls. For example, most descriptive data fields have various constraints built in, and often have a well-developed set of characters from which the entries are chosen. Errors can still occur, however, especially with continuous data where units may be confused (e.g. millimetres and centimetres). Units used should be recorded, and preferably in a separate field as recommended in the SDD standard. Also – standardisation of units within one database should be carried out wherever possible – i.e. agree to use mm throughout, or cm, etc. rather than mix and match which can lead to errors, especially when entry of data is carried out by multiple operators. Tests can be carried out on these fields to look at extremes (e.g. by using cumulative frequency curves as described under the *Spatial Data* above), looking at outliers using Standard Deviations from the mean or median, etc. Often, by graphing the results one can also identify records that are possible errors. Some other error types that may be used to identify errors include (after English 1999).

- ***Missing Data Values***
  Searching for empty fields where values should occur. Where there is need for a "null" or missing value in a field it is good practice to record the reason for the null value in a separate field – for example "not relevant, not measured or unknown"..
- ***Incorrect Data Values***
  This involves searching for typographic errors, transposition of key strokes, data entered in the wrong place (e.g. alphanumeric characters entered into numerical fields), and data values forced into a field that requires a value, but for which the data entry operator doesn't know the value so adds a dummy value. Dummy values are sometimes added into fields to "trick" statistical methods where empty fields or zero values are not allowed. This should be done with care.
- ***Nonatomic Data Values***
  Searching for fields where more than one fact is entered.
- ***Domain Schizophrenia***
  Searching for fields used for purposes for which they may not have been intended.
- ***Duplicate Occurrences***
  Searching for values that may refer to the same real world value. This can occur quite commonly when combining two databases that have used different terminologies.
- ***Inconsistent Data Values***
  Occurs where two related databases may not use the same values lists, and when combined show inconsistencies. This is where the use of transfer standards such as the SDD standard mentioned above come into play.

Documentation of Error

As mentioned in the associated document on *Principles of Data Quality* (Chapman 2005a), documentation of error and error checking is essential to maintain data quality and to avoid duplication of error checking. Without good documentation, users cannot determine the fitness of the data for use.

It is of very little use to anyone if checks of data quality are carried out, and corrections made, if they are not fully documented (Chapman 2005a). A data correction audit trail needs to be maintained as there is always the possibility that perceived errors are not errors at all, and that changes that are made, add new error. Without an audit trail, it may not be possible to undo those "corrections". This is especially important where these checks are being carried out by other than the originator of the data (Chapman 2005a).

There are several ways of developing audit trails (i.e. recording changes made to the database over time as well as recording what data quality control checks have been carried out and when). Audit trails are important so that errors can be recovered, curators and data managers don't carry out checks that have already been carried out, and so that alterations and additions to the data are documented for legal and other purposes (for example, informing users who may have used the data knowing what changes have been made since they last accessed the data). One way of creating audit trails is through the application of a temporal database where a series of time stamps are added, for example a transaction time stamp period during which a fact should be stored in the database (Wikepedia[6]). Another method is to do periodic XML exports of the data of records that have changed, or portions of the data where changes have been made.

As mentioned in Chapman (2005a):

> "*data quality checks carried out on data by a user may identify a number of suspect records. These records may then be checked and found to be perfectly good records and genuine outliers. If this information is not documented in the record, further down the line, someone else may come along and carry out more data quality checks that again identify the same records as suspect*."

Also as mentioned in the associated document on *Principles of Data Quality* (Chapman 2005a):

> *One of the ways of making sure that error is fully documented is to include it in the early planning stages of database design and construction. Additional data quality/accuracy fields can then be incorporated. Fields such as geocode accuracy, source of information for the geocode and elevation, fields for who added the information – was the geocode added by the collector using a GPS, or a data entry operator at a later date using a map at a particular scale, was the elevation automatically generated from a DEM, if so, what was the source of the DEM, its date and scale, etc. All this information will be valuable in later determining whether the information is of value for a particular use or not, and the user of the data can then decide.*

In addition, fields on data validation – the "who, when, how and what" of validation checks carried out should be added to the database to track and audit the validation, error checking and data cleaning carried out on the database. Ideally, these would also be added at the record level as suggested above.

[6] http://en.wikipedia.org/wiki/Temporal_database

Visualisation of Error

There is still a long way to go to develop good error visualisation methods for primary species data. The two requirements of visualisation are

- Visualisation for error checking and cleaning;
- Visualisation for presentation.

The second of these – visualisation for presentation was covered in the associated document on *Principles of Data Quality* (Chapman 2005a).

GIS is the most common method of visualising spatial error for use in checking. Just by mapping primary species data and overlaying it with topographic layers can assist in detecting errors. GIS systems range from simple on-line systems used mostly for on-line mapping and information presentation through to stand-alone systems that vary from the simple to the highly sophisticated.

Many institutions already use GIS for mapping, and these are easily adaptable for use in error checking. Other institutions, however, do not use GIS routinely and consider the purchase of a GIS system beyond their means, but there are free GIS programs available that are easy to learn and simple to use and that will adequately carry out most of the requirements of small collections institutions. At least one of these – Diva-GIS (Hijmans *et al.* 2005) – has been specifically designed for use by small museums and herbaria and includes several error detecting methods described in this document, as well as modelling and visualisation algorithms.

For non-spatial data, one can best visualise error through the use of spreadsheets and graphs. A simple graph of values will quickly identify records that don't fit the patterns. Simple graphs are easy to set up and populate from the database as a standard error checking method.

There is a growing tendency in the spatial community to use techniques such as Monte Carlo Analysis to produce estimates of the likely extent and importance of error (Flowerdew 1991). Monte Carlo analyses lend themselves well to visualisations, and are a good way of conveying error to users. Although some common software that includes Monte Carlo methods have become quite expensive (e.g. Canoco 4.5 for Windows[7] and S-Plus[8]), free alternatives do exist, for example the PopTools add-in for Microsoft Excel (Hood 2005).

Visualising accuracy

As mentioned under *Georeferencing* above, point records of primary specimen records are not really points, but have an error figure associated with them. By mapping the point with its associated accuracy, the "footprint", a good understanding of what the collection actually means, and its relationship to the real world, can be visualised.

This is one area of research that needs urgently pursuing with respect to primary species data – the development of techniques to visualize uncertainty and to show footprints of accuracy. Instead of a collection record being represented as a point of latitude and longitude there is a need to include the accuracy associated with the record and thus present the location as its footprint – a circle, an ellipse, a polygon or even a grid. GIS techniques, such as buffering, provide a good tool for developing footprints such as along rivers or roads. The Biogeomancer program is looking at some aspects of this, but is unlikely to develop a fully operational system in the time available.

[7] http://www.microcomputerpower.com/
[8] http://www.insightful.com/products/splus/default.asp

Cited Tools

1. Software resources

ANUCLIM

Description: A bioclimatic modelling package containing a suite of programs, including the most recent version of BIOCLIM. The program includes a number of methods for identifying errors in the input specimen data.

Version: 5.1 (2004).

Custodian: Centre for Resource and Environmental Studies (CRES), Australian National University, Canberra, Australia.

Cost: $AUD1000.

Reference: Houlder et al. 2000.

Download: http://cres.anu.edu.au/outputs/software.php

BioLink

Description: A software package designed to manage taxon-based information such as nomenclature, distribution, classification, ecology, morphology, illustrations, multimedia and literature.

Version: 2.1 (2005).

Custodian: Australian National Insect Collection, CSIRO, Canberra, Australia.

Cost: Free.

Reference: Shattuck and Fitzsimmons 2000.

Download: http://www.biolink.csiro.au/.

BIOTA

Description: A Biodiversity Data Management system for biodiversity and collections data. Its easy-to-use graphical interface harnesses the power of a fully relational database.

Version: 2.03 (2004).

Custodian: Robert K. Colwell, Connecticut, USA.

Cost: Demo Version Free: Full version: $US200-600.

Reference: Collwell 2002.

Download: http://viceroy.eeb.uconn.edu/Biota2Pages/biota2_download.html

Biótica

Description: Designed to handle curatorial, nomenclatural, geographic, bibliographic and ecological data to assist in the capture and updating.

Version: 4.0 (2003).

Custodian: CONABIO, Mexico City, Mexico.

Cost: $US290.

Reference: Conabio 2002.

Download: http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/distribu_v4.0.html.

BRAHMS

Description: A database software for botanical research and collection management**.** It provides support with the management of names, collection curation and taxonomic research.

Version: 5.58 (2005).

Custodian: University of Oxford, Oxford, UK.

Cost: Free.

Reference: University of Oxford 2004.

Download: http://storage.plants.ox.ac.uk/brahms/defaultNS.html.

Desktop GARP

Description: A software package for prediction and analysis of wild species distributions.
Version: 1.1.3 (2004)
Custodian: University of Kansas, Lawrence, Kansas, USA and Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.
Cost: Free.
Reference: Pereira 2002.
Download: http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1.

Diva-GIS

Description: A geographic information system developed for the analysis of biodiversity data. It includes several simple modelling tools and includes a number of data quality checking algorithms.
Version: 5.0 (2005).
Custodian: R.J. Hijmans *et al.*, University of California, Berkeley.
Cost: Free
Reference: Hijmans *et al.* 2005
Download: http://www.diva-gis.org

eGaz

Description: A program developed to assist museums and herbaria to identify and add geocodes to their specimen records.
Custodian: Australian National Insect Collection, CSIRO, Canberra, Australia.
Cost: Free
Reference: Shattuck 1997.
Download: http://www.biolink.csiro.au/egaz.html

FloraMap

Description: A software tool for predicting the distribution of plants and other organisms in the wild.
Version: 1.02 (2003).
Custodian: Centro Internacional de Agricultura Tropical (CIAT), Columbia.
Cost: $US100.
Reference: Jones and Gladkov 2001.
Download: http://www.floramap-ciat.org/ing/floramap101.htm.

GeoLocate

Description: A georeferencing program to facilitate the task of assigning geographic coordinates to locality data associated with natural history collections.
Version: 2.0 (2003).
Custodian: Tulane Museum of Natural History, Belle Chasse, LA, USA.
Cost: Free.
Reference: Rios and Bart *n.dat*.
Order: http://www.museum.tulane.edu/geolocate/order.aspx.

PATN

Description: A comprehensive and versatile software package for extracting and displaying patterns in multivariate data.
Version: 3.01 (2004).

Custodian: Blatant Fabrications Pty Ltd (Lee Belbin)
Cost: $US299.
Reference: Belbin 2004.
Download: http://www.patn.com.au/.

PopTools

Description: PopTools is a versatile add-in for PC versions of Microsoft Excel that facilitates analysis of matrix population models and simulation and stochastic processes.
Version: 2.6.6 (2005).
Custodian: Greg Hood, Albany, W.A., Australia.
Cost: Free
Reference: Hood 2005
Download: http://www.cse.csiro.au/poptools/.

Specify

Description: A collection management system for natural history museums and herbaria.
Version: 4.6 (2004).
Custodian: Biodiversity Research Center, The University of Kansas, Lawrence, Kansas, USA.
Cost: Free
Reference: University of Kansas 2003a
Download: http://www.specifysoftware.org/Specify/specify/download.

2. On-line resources

BioGeoMancer

Description: A georeferencing service for collectors, curators and users of natural history specimens.
Custodian: Peabody Museum of Natural History, Connecticut, USA.
Reference: Peabody Museum *n.dat.*
Location: http://www.biogeomancer.org
Notes: The BioGeomancer project has recently (2005) been expanded to become a worldwide collaboration of natural history museums with the aim of improving tools for georeferencing and data quality checking. The tools should be available for general use by mid 2006 both as stand-alone products as well as Web Services.

Data Cleaning (CRIA)

Description: An on-line data checking and error identification tool developed by CRIA to help curators of datasets made available via the speciesLink distributed information system to identify possible errors in their databases. Errors include both nomenclatural and geographic.
Custodian: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.
Location: http://splink.cria.org.br/dc
Notes: Some of the algorithms developed in this tool (especially the geographic tools) are being incorporated into the BioGeomancer toolkit as part of a worldwide collaborative project due for completion in mid 2006.

geoLoc

Description: A tool to assist biological collections in georeferencing their data.
Custodian: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.
Location: http://splink.cria.org.br/geoloc?&setlang=en

Georeferencing Calculator

Description: A java applet created to aid in the georeferencing of descriptive localities such as found in museum-based natural history collections.
Custodian: University of California, Berkeley, CA, USA.
Location: http://manisnet.org/manis/gc.html

Lifemapper

Description: Screensaver software that uses the Internet to retrieve records of plants and animals from natural history museums and uses modelling algorithms to predict distributions.
Custodian: Biodiversity Research Center, The University of Kansas, Lawrence, Kansas, USA.
Location: http://www.lifemapper.org/

spOutlier

Description: An automated tool used to detect outliers in latitude, longitude and altitude, and to identify errant on-shore or off-shore records in natural history collections data.
Custodian: Centro de Referência em Informação Ambiental (CRIA), Campinas, Brazil.
Location: http://splink.cria.org.br/outlier?&setlang=en

3. Standards and Guidelines

DELTA

Description: The DELTA format (DEscription Language for TAxonomy) is a flexible method for encoding taxonomic descriptions for computer processing.
Standard: Adopted by TDWG as a standard for data exchange.
Reference: Dallwitz *et al.* 1993.
Location: http://biodiversity.uno.edu/delta/

HISPID

Description: Herbarium Information Standards and Protocols for Interchange of Data.
Custodian: Committee of Heads of Australian Herbaria. Adopted as a TDWG Standard.
Reference: Conn 1996, 2000.
Location: http://plantnet.rbgsyd.nsw.gov.au/Hispid4/

MaNIS Georeferencing Guidelines

Description: Contains information about assigning geographic coordinates, and maximum error distances for those coordinates, to locality descriptions.
Custodian: University of California, Berkeley, CA, USA.
Location: http://manisnet.org/manis/GeorefGuide.html.

Manual de Procedimentos para Georreferenciar

Description: Manual developed by CONABIO in Mexico as guidelines for georeferencing natural history collections. In Spanish with an English abstract being prepared.
Reference: CONABIO 2005.
Location: Not yet available electronically.

MaPSTeDI Georeferencing Guidelines

Description: Guide to the specimen georeferencing process in the MaPSTeDI project.
Custodian: University of Colorado Regents, Denver, CO, USA.
Location: http://mapstedi.colorado.edu/geocoding.html

Plant Names in Botanical Databases

Description: The purpose of this standard is to specify how scientific names of plants may be organised in botanical databases.
Custodian: Taxonomic Databases Working Group (TDWG)..
Location: http://www.tdwg.org/plants.html

SDD

Description: The SDD subgroup of TDWG was established to develop an international XML-based standard for capturing and managing descriptive data for organisms.
Custodian: Taxonomic Databases Working Group (TDWG)..
Location: http://160.45.63.11/Projects/TDWG-SDD/index.html

TDWG Standards

Description: The Taxonomic Databases Working Group (TDWG) has been developing standards for use with biodiversity data for many years. Standards have been, and are being developed for a range of issues related to the storage, documentation, and distribution of species and species-occurrence data
Custodian: Taxonomic Databases Working Group (TDWG).
Location: http://www.tdwg.org/standrds.html
http://www.tdwg.org/subgrops.html

Conclusion

Errores ad sua principia referre, est refellere
To refer errors to their origin is to refute them.
(Ref. 3 Co. Inst. 15)

The information age has meant that collections' institutions have become an integral part of the environmental decision making process and politicians are increasingly seeking relevance and value in return for the resources that they put into those institutions. It is thus in the best interests of collections' institutions that they produce a quality product if they are to continue to be seen as a value-adding resource by those supplying the funding.

Best practice for databased information in museums and herbaria and institutions maintaining survey and observational information means making the data as accurate and possible, and using the most appropriate techniques and methodologies to ensure that the data are the best they can possibly be. To ensure that this is the case, it is essential that data entry errors are reduced to a minimum, and that on-going data cleaning and validation are integrated into day-to-day data and information management protocols.

There is no such thing as good quality data or bad quality data (Chapman 2005a). Data are data, and their use will determine their quality. Nevertheless, data providers need to ensure that the data are as free from error as it is possible to make them. No one test alone will ever be sufficient to identify all errors in a dataset, and thus it is important to use a combination of methods that best fit the circumstances of the organisation using them and the data contained therein. In addition collaboration between institutions, data providers, scientists and IT professionals as well as the users of the data is needed to improve data quality not only within individual collection institutions, but also across the totality of collections as combination takes place.

Perhaps the most important data management practice is good documentation. No matter what tests have been carried out on the data, they should be fully documented. Only in this way, can users of the data be truly informed as to their nature and likely accuracy.

In this period of increasing data and information exchange, the reputation of a collections' institution is likely to hinge on the quality and availability of its information (Redman 1996, Dalcin 2004), rather than on the quality of its scientists, as has been the case in the past. This is a fact of life, and the two can no longer be separated. Good data and information management must run side by side with good science and together they should lead to good data and information.

Acknowledgements

Many colleagues and organisations around the world have contributed to this paper either directly or indirectly. The ideas elaborated, could not have been developed without discussions with like-minded individuals, and this has been the case over a period of 30 years. Only by making our knowledge available to others so that they can build on that knowledge and move it another step forward can our science progress. In recent years there has been a massive increase in willingness to share information prior to publication, and in developing collaborative projects across both social and political boundaries. The introduction of the Internet has made this so much easier to do and it is the only way that our discipline can keep up with the massive progress in data and information services that are needed.

In particular, I would like to make particular 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, the Peabody Museum in Connecticut, and the University of California in Berkeley in particular, has made this task possible. In addition, organisations like Conabio in Mexico, the University of Kansas, CSIRO in Australia and the University of Colorado, as well as others too numerous to mention, have helped bring us to the stage we are today in primary species data quality management. I thank them for their unique ideas and constructive criticism.

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 editing and review 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; Mervyn Mansell, USDA-Aphis, Pretoria, South Africa; Francisco Pando, Real Jardín Botánico, Madrid, Spain; 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 managing its review and editing.

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:

Armstrong, J.A. 1992. The funding base for Australian biological collections. *Australian Biologist* **5(1):** 80-88.

ABRS. 2004. *Australian Faunal Directory*. Canberra: Australian Biological Resources Study. http://www.deh.gov.au/biodiversity/abrs/online-resources/abif/fauna/afd/index.html [Accessed 12 Apr. 2005].

ANBG. 2003. *Australian Plant Name Index*. Canberra: Australian National Botanic Gardens. http://www.anbg.gov.au/apni/index.html [Accessed 12 Apr. 2005].

Barnett, V. and Lewis, T. 1994. *Outliers in Statistical Data*. Chichester, UK: Wiley and Sons.

Beaman, R.S. 2002. Automated georeferencing web services for natural history collections **in** *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* http://www.cria.org.br/eventos/tdbi/flora/reed [Accessed 12 Apr. 2005]

Beaman, R., Wieczorek, J. and Blum, S. 2004. Determining Space from Place for Natural History Collections in a Distributed Library Environment. *D-Lib Magazine* Vol. 10(5). http://www.dlib.org/dlib/may04/beaman/05beaman.html [Accessed 12 Apr. 2005].

Belbin, L. 2004. *PATN vers. 3.01*. Blatant Fabrications http://www.patn.com.au [Accessed 12 Apr. 2005].

Berendsohn, W.G. 1997. A taxonomic information model for botanical databases: the IOPI model. *Taxon* 46: 283-309.

Berendsohn, W., Güntsch, A. and Röpert, D. (2003). *Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections.* Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Members/irc/gbif/digit/library?l=/digitization_collections/contract_2003_report/
[Accessed 13 Apr. 2005].

BioCASE. 2003. *Biological Collection Access Service for Europe*. http://www.biocase.org [Accessed 12 Apr. 2005].

Birds Australia. 2004. *Birds Australia Rarities Committee (BARC).* http://users.bigpond.net.au/palliser/barc/barc-home.html [Accessed 12 Apr. 2005].

Bisby, F.A. 1994. *Plant Names in Botanical databases.* TDWG Standard. http://www.tdwg.org/plants.html [Accessed 12 Apr. 2005].

Bisby, F.A., Zarucchi, J.L., Schrire, B.L., Roskov, Y.R., Heald, J. and White, R.J. (eds). 2002. *ILDIS World Database of Legumes* ver. 6.05. http://www.ildis.org/ [Accessed 12 Apr. 2005].

Blakers, M., Davies, S.J.J.F. and Reilly, P.N. 1984. *The Atlas of Australian Birds*. Melbourne: Melbourne University Press.

Blum, S. 2001. *Georeferencing Natural History Collection Localities at the California Academy of Sciences*. http://www.calacademy.org/research/informatics/GeoRef/index.html [Accessed 12 Apr. 2005].

Brickell, C.D., Baum, B.R., Hetterscheid, W.L.A., Leslie, A.C., McNeill, J., Trehane, P., Vrugtman, F. and Wiersema, J.H. (eds) 2004. *International Code for Cultivated Plants* ed. 7. Edinburgh, U.K.: ISHS. http://www.actahort.org/books/647/ [Accessed 11 Apr. 2005].

Brummitt, R.K. and Powell, C.E. (eds). 1992 *Authors of Plant Names.* Kew: Royal Botanic Gardens, Kew. http://www.ipni.org/index.html [Accessed 12 Apr. 2005]

Burbidge, A.A. 1991. Cost Constraints on Surveys for Nature Conservation **in** Margules, C.R. and Austin, M.P. (eds). *Nature Conservation: Cost Effective Biological Surveys and Data Analysis.* Canberra: CSIRO.

Burrough, P.A. and McDonnell, R.A. 1998. *Principals of Geographical Information Systems.* Oxford, UK: Oxford University Press.

Busby, J.R. 1991. BIOCLIM – a bioclimatic analysis and prediction system. Pp. 4-68 **in** Margules, C.R. and Austin, M.P. (eds) *Nature Conservation: Cost Effective Biological Surveys and data Analysis*. Melbourne: CSIRO.

Chapman, A.D. 1988. Karl Domin in Australia **in** *Botanical History Symposium. Development of Systematic Botany in Australasia. Ormond College, University of Melbourne. May 25-27, 1988.* Melbourne: Australian Systematic Botany Society, Inc.

Chapman, A.D. 1991. Australian Plant Name Index pp. 1-3053. *Australian Flora and Fauna Series* Nos 12-15. Canberra: AGPS.

Chapman, A.D. 1992. Quality Control and Validation of Environmental Resource Data **in** *Data Quality and Standards: Proceedings of a Seminar Organised by the Commonwealth Land Information Forum, Canberra, 5 December 1991*. Canberra: Commonwealth land Information Forum.

Chapman, A.D. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data pp. 409-418 **in** Lowell, K. and Jaton, A. eds. *Spatial accuracy assessment: Land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press.

Chapman, A.D. *et al*. 2002. *Guidelines on Biological Nomenclature*. Canberra: Environment Australia. http://www.deh.gov.au/erin/documentation/nomenclature.html [Accessed 12 Apr. 2005].

Chapman, A.D. 2004. Guidelines on Biological Nomenclature. Brazil edition. Appendix J to *Sistema de Informação Distribuído para Coleções Biológicas: A Integração do Species Analyst e SinBiota. FAPESP/Biota process no. 2001/02175-5 March 2003 – March 2004.* Campinas, Brazil: CRIA 11 pp. http://splink.cria.org.br/docs/appendix_j.pdf [Accessed 12 Apr. 2005].

Chapman, A.D. 2005a. *Principles of Data Quality*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/data_quality [Accessed 1 Aug. 2005].

Chapman, A.D. 2005b. *Uses of Primary Species-Occurrence Data*. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/uses_of_data [Accessed 1 Aug. 2005].

Chapman, A.D. and Busby, J.R. 1994. Linking plant species information to continental biodiversity inventory, climate and environmental monitoring 177-195 **in** Miller, R.I. (ed.). *Mapping the Diversity of Nature*. London: Chapman and Hall.

Chapman, A.D. and Milne, D.J. 1998. *The Impact of Global Warming on the Distribution of Selected Australian Plant and Animal Species in relation to Soils and Vegetation*. Canberra: Environment Australia

Chapman, A.D., Bennett, S., Bossard, K., Rosling, T., Tranter, J. and Kaye, P. 2001. Environment Protection and Biodiversity Conservation Act, 1999 – Information System. *Proceedings of the 17th Annual Meeting of the Taxonomic Databases Working Group, Sydney, Australia 9-11 November 2001*. Powerpoint: http://www.tdwg.org/2001meet/ArthurChapman_files/frame.htm [Accessed 12 Apr. 2005].

CHAH 2002. *AVH - Australian's Virtual Herbarium*. Australia: Council of Heads of Australian Herbaria. http://www.chah.gov.au/avh/avh.html [Accessed 12 Apr. 2005].

Chrisman, N.R., 1991. The Error Component in Spatial Data. pp. 165-174 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Christidis, L. & Boles, W.E. 1994. *Taxonomy and Species of Birds of Australia and its Territories*. Royal Australasian Ornithologists Union, Melbourne. 112 pp.

Clarke, K.C. 2002. *Getting Started with Geographic Information Systems*, 4th edn. Upper Saddle River, NJ, USA: Prentice Hall. 352 pp.

Colwell, R.K. 2002. *Biota: The Biodiversity Database Manager*. Connecticut, USA: University of Connecticut http://viceroy.eeb.uconn.edu/Biota [Accessed 12 Apr. 2005].

CONABIO. 2002. *The Biótica Information system*. Mexico City: Comisión national para el conocimiento y uso de la biodiversidad. http://www.conabio.gob.mx/informacion/biotica_ingles/doctos/acerca_biotica.html [Accessed 12 Apr. 2005

CONABIO. 2005. *Manual de Procedimentos para Georreferenciar*. Mexico: Comisión para el Conocimiento y Uso de la Biodiversidad México (CONABIO).

Conn, B.J. (ed.) 1996. *HISPID3. Herbarium Information Standards and Protocols for Interchange of Data*. Version 3 (Draft 1.4). Sydney: Royal Botanic Gardens. http://www.bgbm.org/TDWG/acc/hispid30draft.doc [Accessed 10 Apr 2005]

Conn, B.J. (ed.) 2000. *HISPID4. Herbarium Information Standards and Protocols for Interchange of Data*. Version 4 – Internet only version. Sydney: Royal Botanic Gardens. http://plantnet.rbgsyd.nsw.gov.au/Hispid4/ [Accessed 30 Jul. 2003].

Croft, J.R. (ed.) 1989. *HISPID – Herbarium Information Standards and Protocols for Interchange of Data*. Canberra: Australian National Botanic Gardens.

CRIA. 2002. *speciesLink*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/ [accessed 12 Apr. 2005]

CRIA. 2004a. *GeoLoc-CRIA*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [Accessed 12 Apr. 2005].

CRIA. 2004b. *spOutlier-CRIA*. Centro de Referência em Informação Ambiental. http://splink.cria.org.br/tools/ [accessed 1 Mar. 2005]

CRIA (2005), *speciesLink. Dados e ferramentos. Data Cleaning*. Centro de Referência em Informação Ambiental. http://splink.cria.org.br/dc/ [Accessed 12 Apr. 2005].

Dalcin, E.C. 2004. *Data Quality Concepts and Techniques Applied to Taxonomic Databases*. Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southampton. November 2004. 266 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf [Accessed 7 Jan. 2005].

Dallwitz, M.J., Paine, T.A. and Zurcher, E.J. (1993). *User's guide to the DELTA System: a general system for processing taxonomic descriptions*. 4th edn. http://delta-intkey.com/ [Accessed 12 Apr. 2005].

DEH. 2005a. *Threatened Species*. Canberra: Department of Environment and Heritage. http://www.deh.gov.au/biodiversity/threatened/species/index.html [Accessed 12 Apr. 2005].

DEH. 2005b. *Species Profile and Threats Database*. Canberra : Department of Environment and Heritage. http://www.deh.gov.au/cgi-bin/sprat/public/sprat.pl [Accessed 7 Apr. 2005].

Dorr, LJ. 1997. *Plant Collectors in Madagascar and the Comoro Islands*. Kew, UK: Royal Botanic Gardens, Kew.

English, L.P. 1999. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., New York.

ESRI. 2003. *ArcSDE: The GIS Gateway to Relational Databases*. http://www.esri.com/software/arcgis/arcinfo/arcsde/overview.html [Accessed 12 Apr. 2005

Farr, E. and Zijlstra, G. (eds). *n.dat. Index Nominum Genericorum (Plantarum)*. On-line version. http://ravenel.si.edu/botany/ing/ [Accessed 21 Jul. 2004].

Flowerdew, R., 1991. Spatial Data Integration. pp. 375-387 **in**: Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

Froese, R. and Bisby, F.A. (eds). 2004. *Catalogue of Life 2004*. Los Baňos, Philippines: Species 2000. http://www.sp2000.org/AnnualChecklist.html [Accessed 7 Apr. 2005].

Froese, R. and Pauly, D. 2004. *Fishbase*. Ver. 05/2004. The Philippines: World Fish Center. http://www.fishbase.org/ [Accessed 10 Apr. 2005].

Fundación Biodiversidad 2005. *Proyeto Anthos – Sistema de información sobre los plantas de España*. http://www.programanthos.org/ [Accessed 8 Apr. 2005].

Gatrell, A.C. 1991. Concepts of Space and Geographical Data. pp. 119-134 **in:** Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) *Geographical Information Systems* Vol. 1, Principals: Longman Scientific and Technical.

GBIF. 2003a. *What is GBIF?* http://www.gbif.org/GBIF_org/what_is_gbif [Accessed 12 Apr. 2005].

GBIF. 2003b. *GBIF Work Program 2004*. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.org/GBIF_org/wp/wp2004/GB7_20WP2004-v1.0-approved.pdf [Accessed 12 Apr. 2005].

GBIF. 2004. *Data Portal*. Copenhagen: Global Biodiversity Information Facility. http://www.gbif.net/portal/index.jsp. [Accessed 12 Apr. 2005].

Geographic Names Board. 2003. *Guidelines for Naming of Roads*. Sydney: Geographic Names Board of New South Wales. http://www.gnb.nsw.gov.au/newsroom/road_naming_guideline.pdf [Accessed 21 Jul. 2004].

Greuter, W. *et al.* 1984-1989. *Med-Checklist: a critical inventory of vascular plants of the circum-mediterranean countries*. 4 Vols. Botanical Garden and Botanical Museum Berlin-Dahlem.

Hepper, F.N. and Neate, F. 1971. *Plant Collectors in West Africa*. Utrecht, The Netherlands: Oosthoeks' Uitgeversmaatschappij.

Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrentes, I. and Rojas, E. 2005 *DIVA-GIS Version 5. A geographic information system for the analysis of biodiversity data*. http://www.diva-gis.org [Accessed 30 Jul. 2004].

Hobern, D. and Saarenmaa, H. 2005. *GBIF Data Portal Strategy*. Draft Version 0.14. Copenhagen, Denmark: Global Biodiversity Information Facility. http://circa.gbif.net/Public/irc/gbif/dadi/library?l=/architecture/portal_strategy_1/ [Accessed 7 Apr. 2005].

Hood, G.M. 2005. *PopTools version 2.6.6*. Canberra: CSIRO Sustainable Ecosystems. http://www.cse.csiro.au/poptools [Accessed 13 Apr. 2005].

Houlder, D. Hutchinson, M.J., Nix, H.A. and McMahaon, J. 2000. *ANUCLIM 5.1 Users Guide*. Canberra: Cres, ANU. http://cres.anu.edu.au/outputs/anuclim.php [Accessed 12 Apr. 2005].

IAPT. 1997. *Names in Current Use for Extant Plant Genera ver. 1.0*. on-line version. International Association for Plant Taxonomy. http://www.bgbm.org/iapt/ncu/genera/Default.htm [Accessed 12 Apr. 2005].

ICSM. 2001. *Guidelines for the Consistent Use of Place Names*. Intergovernmental Committee on Survey and Mapping: Committee for Geographic Names in Australia. http://www.icsm.gov.au/icsm/cgna/consistent_pnames.pdf. [Accessed 12 Apr. 2005].

Index Herbariorum. (1954-1988) *Index Herbariorum Part 2: Collectors*. Various compilers. Utrecht/Antwerp, The Hague/Boston
Part 2(1): Collectors A-D (1954). *Regnum Vegetabile* vol. 2 (A-D),
Part 2(2): Collectors E-H (1957). *Regnum Vegetabile* vol. 9 (E-H),
Part 2(3): Collectors I-L (1972). *Regnum Vegetabile* vol. 86 (I-L),
Part 2(4): Collectors M (1976). *Regnum Vegetabile* vol. 93 (M),
Part 2(5): Collectors N-R (1983). *Regnum Vegetabile* vol. 189 (N-R),
Part 2(6): Collectors S (1986). *Regnum Vegetabile* vol. 114 (S),
Part 2(7): Collectors T-Z (1988). *Regnum Vegetabile* vol. 117 (T-Z).

IPNI. 1999. *International Plant Names Index*. http://www.ipni.org/index.html [accessed 12 Apr. 2005].

IOPI. 2003. *Global Plant Checklist*. International Organization for Plant Information (IOPI). http://www.bgbm.fu-berlin.de/IOPI/GPC/ [Accessed 12 Apr. 2005].

Johnson, R.A. and Wichern, D.W. 1998. *Applied Multivariate Statistical Analysis*. 4th edn. New York, NY: Prentice Hall.

Jones P.G. and Gladkov, A. 2001. *Floramap Version 1.01*. Cali, Colombia: CIAT. http://www.floramap-ciat.org/ing/floramap101.htm [Accessed 12 Apr. 2005].

Koch, I. 2003. *Coletores de plantas brasileiras*. Campinas: Centro de Referência em Informação Ambiental. http://splink.cria.org.br/collectors_db [Accessed 12 Apr. 2005].

Lampe, K.-H. and Riede, K. 2002. *Mapping the collectors: the georeferencing bottleneck*. Poster given to TDWG meeting, Indaiatuba, Brazil. http://www.cria.org.br/eventos/tdbi/bis/Poster-200dpi.html [accessed 12 Apr. 2005].

Legendre P. and Legendre L. (1998): Numerical Ecology. *Developments in Environmental Modeling* 20, Second English Edition, Elsevier, Amsterdam, 853p. http://www.bio.umontreal.ca/legendre/numecol.html [Accesssed 12 Apr. 2005].

Lindemeyer, D.B., Nix, H.A., McMahon, J.P., Hutchinson, M.F. and Tanton, M.T. 1991. The Conservation of Leadbeater's Possum, *Gymnobelidus leadbeateri* (McCoy): A Case Study of the Use of Bioclimatic Modelling. *J. Biogeog.* **18:** 371-383.

Lowry, R. 2005. Concepts and Applications of Inferential Statistics. *VassarStats: Web Site for Statistical Computation*. http://faculty.vassar.edu/lowry/webtext.html [Accessed 8 Apr. 2005]

Maletic, J.I. and Marcus, A. 2000. Data Cleansing: Beyond Integrity Analysis pp. 200-209 in *Proceedings of the Conference on Information Quality (IQ2000)*. Boston: Massachusetts Institute of Technology. http://www.cs.wayne.edu/~amarcus/papers/IQ2000.pdf [Accessed 12 Apr. 2005].

MaNIS. 2001. *The Mammal Networked Information System*. http://manisnet.org/manis [Accessed 12 Apr. 2005].

Marcus, A., Maletic, J.I. and Lin, K.-I. 2001. Ordinal Association Rules for Error Identification in Data Sets pp. 589-591 in *Proceedings of the 10th ACM Conference on Information and Knowledge Management (ACM CIKM 2001). Atlanta, GA.* http://www.cs.wayne.edu/~amarcus/papers/cikm01.pdf [Accessed 12 Apr. 2005].

Margules, C.R. and Redhead, T.D. 1995. *BioRap. Guidelines for using the BioRap Methodology and Tools*. Canberra: CSIRO. 70pp.

Marino, A., Pavarin, F., de Souza, S. and Chapman, A.D. in prep. *Simple on line tools for geocoding and validating biological data*. To be submitted.

Neldner, V.J., Crossley, D.C. and Cofinas, M. 1995. Using Geographic Information Systems (GIS) to Determine the Adequacy of Sampling in Vegetation Surveys. *Biological Conservation* 73: 1-17.

Nix, H.A. 1986. A biogeographic analysis of Australian elapid snakes in Longmore, R.C. (ed). Atlas of Australian elapid snakes. *Australian Flora and Fauna Series* No. **7:** 4-15. Canberra: Australian Government Publishing Service.

NMNH. 1993. *RapidMap. Geocoding locality descriptions associated with herbarium specimens*. U.S. National Musuem of Natural History and Bernice P. Bishop Museum, Honolulu. http://users.ca.astound.net/specht/rm/ [Accessed 12 Apr. 2005].

Peabody Museum. *n.dat*. *BioGeoMancer*. http://www.biogeomancer.org [Accessed 12 Apr. 2005].

Peterson, A.T., Navarro-Siguenza, A.G. and Benitez-Diaz, H. 1998. The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. *Ibis* **140:** 288-294.

Peterson, A.T., Stockwell, D.R.B. and Kluza, D.A. 2002. Distributional Prediction Based on Ecological Niche Modelling of Primary Occurrence Data pp. 617-623 **in** Scott, M.J. *et al.* eds. *Predicting Species Occurrences. Issues of Accuracy and Scale*. Washington: Island Press.

Peterson, A.T., Navarro-Siguenza, A.G. and Pereira, R.S. 2003. Detecting errors in biodiversity data based on collectors' itineraries. *Bulletin of the British Ornithologists' Club* 124: 143-151. http://www.specifysoftware.org/Informatics/bios/biostownpeterson/PNP_BBOC_2004.pdf. [Accessed 12 Apr. 2005].

Pfeiffer, U., Poersch, T. and Fuhr, N. 1996. Retrieval effectiveness of proper name search methods. *Information Processing and Management* **32(6):** 667-679.

Platnick, N.I. 2004. *The World Spider Catalog*. New York: The American Museum of Natural History. http://research.amnh.org/entomology/spiders/catalog81-87/INTRO3.html [Accessed 12 Apr. 2005].

Podolsky, R. 1996. *Software Tools for the Management and Visualization of Biodiversity Data*. NY, USA: United Nations Development Project. http://www3.undp.org/biod/bio.html [Accessed 13 Apr. 2005].

Pollock, J.J. and Zamora, A. 1984. Automatic spelling correction in scientific and scholarly text. *Communications of ACM* **27(4):** 358-368.

Redman, T.C. 1996. *Data Quality for the Information Age*. Artech House Inc.

Redman, T.C. 2001. *Data Quality: The Field Guide*. Boston, MA: Digital Press.

Rios, N.E. and Bart, H.L. Jr. *n.dat. GEOLocate. Georeferencing Software. User's Manual*. Belle Chasse, LA, USA: Tulane Museum of Natural History. http://www.museum.tulane.edu/geolocate/support/manual_ver2_0.pdf [Accessed 12 Apr. 2005].

Roughton, K.G. and Tyckoson, D.A. 1985. Brousing with sound: Sound-based codes and automated authority control. *Information Technology and Libraries* **4(2):**130-136.

Ruggiero, M. (ed.) 2001. *Integrated Taxonomic Information System*. http://www.itis.usda.gov/ [Accessed 12 Apr. 2005].

Pereira, R.S. 2002. *Desktop Garp*. Lawrence, Kansas: University of Kansas Center for Research. http://www.lifemapper.org/desktopgarp/Default.asp?Item=2&Lang=1 [Accessed 13 Apr. 2005].

Shattuck, S.O. 1997. eGaz, The Electronic Gazetteer. *ANIC News* **11:** 9 http://www.ento.csiro.au/biolink/egaz.html [Accessed 12 Apr. 2005].

Shattuck, S.O. and Fitzsimmons, N. 2000. *BioLink, The Biodiversity Information Management System*. Melbourne, Australia: CSIRO Publishing. http://www.ento.csiro.au/biolink/software.html [Accessed 12 Apr. 2005].

Steenis-Kruseman, M.J. van 1950. Malaysian Plant Collectors and Collections. *Flora Malesiana* Vol. 1. Leiden, The Netherlands.

Stockwell, D. and Peters, D. 1999. "The GARP modelling system: problems and solutions to automated spatial prediction." *International Journal of Geographical Information Science* **13**(2): 143-158.

University of Colorado Regents. 2003a. *mapstedi. Geocoding*. Denver: University of Colorado MaPSTeDI project. http://mapstedi.colorado.edu/geocoding.html [Accessed 12 Apr. 2005].

University of Colorado Regents. 2003b. *GeoMuse*. Denver: University of Colorado MaPSTeDI project. http://www.geomuse.org/mapstedi/client/start.jsp [Accessed 12 Apr. 2005].

University of Kansas. 2003a. *Specify*. Biological Collections Management. Lawrence, Kansas: University of Kansas http://www.specifysoftware.org/Specify/ [Accessed 12 Apr. 2005].

University of Kansas. 2003b. *LifeMapper*. Lawrence, Kansas: University of Kansas – Informatics Biodiversity Research Center. http://www.lifemapper.org/ [Accessed 12 Apr. 2005

University of Oxford. 2004. *BRAHMS. Botanical Research and Herbarium Management System*. Oxford, UK: University of Oxford http://storage.plants.ox.ac.uk/brahms/defaultNS.html [Accessed 27 Jul 2004].

Weber, W.A. 1995. Vernacular Names: Why Oh Why?. *Botanical Electrical News* No. 109. http://www.ou.edu/cas/botany-micro/ben/ben109.html [Accessed 7 Apr. 2005].

Weiher, E. and Keddy, P. (eds). 1999. *Ecological Assembly Rules: Perspectives, Advances, Retreats*. Cambridge, UK: Cambridge University Press. 418 pp.

Wieczorek, J. 2001a. *MaNIS: Georeferencing Guidelines*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/GeorefGuide.html [Accessed 12 Apr. 2005].

Wieczorek, J. 2001b. *MaNIS: Georeferencing Calculator*. Berkeley: University of California, Berkeley - MaNIS http://manisnet.org/manis/gc.html [Accessed 12 Apr. 2005].

Wieczorek, J. and Beaman, R.S. 2002. Georeferencing: Collaboration and Automation in *Symposium: Trends and Developments in Biodiversity Informatics, Indaiatuba, Brazil 2002* http://www.cria.org.br/eventos/tdbi/bis/georeferencing [Accessed 12 Apr. 2005].

Wieczorek, J., Guo, Q. and Hijmans, R.J. 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science* 18(8): 745-767.

Wiley, E.O. 1981. *Phylogenetics: the theory and practice of phylogenetic systematics*. New York: John Wiley & Sons.

Williams, P.H., Marguiles, C.R. and Hilbert, D.W. 2002. Data requirements and data. sources for biodiversity priority area selection*. J. Biosc.* **27(4):** 327-338.

Index