Automated geoparsing and georeferencing of Malesian collection locality data

Reed S. Beaman and Barry J. Conn

Abstract

Beaman, Reed S. (Natural History Museum and Biodiversity Research Center, University of Kansas, 1345 Jayhawk Boulevard, Lawrence, KS 66045, USA) and Conn, Barry J. (Royal Botanic Gardens Sydney, Mrs Macquaries Road, Sydney, NSW 2000 Australia) 2003. Geoparsing and georeferencing of Malesian collection locality data. Telopea *10(1) 43–52.* Some form of geographic reference is almost always present on specimen labels, an essential source of information for mapping species distributions and performing biogeographic analyses. The prospect of databasing large herbarium collections is now reality, but the task of manually georeferencing each specimen would be enormous. The fields of biological informatics and geomatics (biogeomatics) provide tools that streamline and automate acquisition, sharing, analysis, and visualisation of biogeographic data. Digitisation of specimens, particularly type specimens is now commonplace, but specimen imaging and optical character recognition (OCR) may also facilitate the data entry process. Natural language processing of digital data significantly reduces the time required to database and georeference a specimen. A prototype for a geoparsing and georeferencing web service has been developed that utilises a digital gazetteer of over 330 000 Malesian place names. This service is demonstrated using Urticaceae collections from Malesia and comparisons are made between automated and manual georeferencing methods.

Introduction

Biological collections document the diversity and distribution of life on earth. These data are now becoming accessible in networked databases for research, conservation, environmental management and educational uses. Though each discipline has specific data capture needs, every natural history collection shares the problem of transforming descriptive locality information associated with specimens into quantitative spatial values. For most of the last 300 years, specimen locality information has been recorded as cardinal offsets to political or geographical features (eg. '24 km N of Springfield, along a stream bed'). These data can be particularly time-consuming to quantify since they require interpretation and additional resources such as maps and gazetteers in order to be recorded in a standard geographic format (that is, to be georeferenced). Quantitative georeferences for biological specimen localities are the critical data that, in association with the taxon identity and the collection date, allow researchers to visualise and model the known and potential distribution of taxa. This particular area of biological data digitisation offers not only the greatest challenge, but also the greatest potential for a solution that provides significant cost savings for the entire biological collections community. For biological collections, a georeferencing solution must be a cohesive, interoperable system that provides natural language processing, geospatial integration, spatial error analysis, and validation based on species-specific geographic distributions. This paper summarises the development of an automated georeferencing tool for biological collections.

Background

The biological collections held by museums and herbaria worldwide document the biological exploration of the planet and are the primary research archives of biotic diversity. These collections provide the basis for the identification, description, comparison of taxa, and documentation of the occurrence of taxa in space and time. To illustrate the scale of information stored in such collections, there are approximately 6 million botanical collections held by the herbaria associated with the *Council of Heads of Australian Herbaria* (CHAH: Anon. 2001), 500 million specimens of animals and plants in museums and herbaria of the United States of America, and an estimated 2.5 billion natural history specimens in collections worldwide (Duckworth et al. 1993). Specimen data document the identities, habitats, histories, and spatial distributions of the 1.5–8 million described species, and provide the fundamental resource for identifying the estimated tens of millions of species that remain to be discovered and described (Wilson 2000). These essential records are the knowledge foundation for such diverse disciplines as biological systematics, environmental planning, conservation, genetic engineering, and medicine.

HISCOM, the *Herbarium Information Systems Committee* (Conn & Brooks 1998) is an Australian national association of Federal and State/Territory herbaria whose mandate is to develop infrastructure and to complete the process of making herbarium specimen data available to CHAH herbaria and the general public. The HISCOM partners have long been active globally in building interchange standards for botanical collections data (Conn 1996, 2000). The *Council of Heads of Australian Herbaria* currently have funding through to the year 2005 for digital library development as part of the *Australia's Virtual Herbarium* project (AVH), a consortium of Federal and State/Territory agencies in Australia (Barker 1998). The AVH digital library consists of four fundamental strategies; namely, (i) the building, sharing, and preservation of digital collections; (ii) creation of tools (particularly, identification tools) and services; (iii) influencing and supporting innovation in communication between users; and (iv) the development of strategic partnerships for further digital library development. However, the primary deliverables for this funding are the cataloguing of approximately 6.5 million herbarium specimens in Australia and establishing interoperability among the partners to reduce duplication of effort.

Recent developments in distributed database networks (eg. *Australia's Virtual Herbarium* – Barker 1998 and *The Species Analyst* – Anon. 1998) have begun to provide widespread access to biological collections data. However, the AVH is not currently interoperable with U.S.-based digital specimen libraries, such as *The Species Analyst* (Anon. 1998) and *Lifemapper* (Anon. 2002). There is presently a significant duplication of effort between the Australian and U.S. initiatives, and increased collaboration and interoperability are desirable for all parties. In addition, despite the recognised importance of these data, many collections remain largely inaccessible because they are not fully digitised. Internationally, herbaria and museums are participating in the painstaking process of digitally capturing specimen data. Tools to make this process more efficient are desperately needed. Georeferencing and validation are services from which the entire biodiversity community can benefit considerably.

The need for georeferencing of biological collections

Biodiversity is inextricably linked to geography. Common among the data for all biological collections is locality information. It is exactly this common thread of information that forms the basis for the investigation not only of individual species,

but also of entire ecological communities. Maier et al. (2000) concluded that the single most important factor influencing work in biodiversity and ecosystem informatics is the problem of complexity, including georeferencing and species referencing. Georeferencing provides the means to link specimen data to the rapidly growing body of spatial environmental data for interdisciplinary research into complex phenomena.

Since most biological collection locality data are written in the form of descriptive localities (often with cardinal offsets), these data do not easily lend themselves to spatial filtering, comparison, or analysis. Locality data interpreted as spatial coordinates with associated measures of uncertainty (accuracy and precision) are more readily queried. Furthermore, the results of such georeferenced data queries are much more readily applied. The importance of capturing localities as spatial data (georeferencing) is recognised as a priority by the 'Digitisation of Natural History Collection Data' Subcommittee of GBIF (GBIF 2002). While it must be remembered that there are limits to the scalability of such data (because collectors were frequently not mindful of computers, analysis or even basic mapping, when they recorded collection information), the development of a method for rapidly digitising locality information in a standard useful form will make a major contribution not only to the digitisation of biological collections worldwide, but also to the vast scientific and public communities that rely on collections information.

Approximately 70% of the herbarium collections held in the *NSW Collections* database of the Royal Botanic Gardens Sydney (NSW) lack georeferences. To illustrate the general lack of georeferenced collections, a sample of 16,300 records from the specialist rapid-entry database (*Rapid*) used by data-processing staff at NSW was analysed (Fig. 1). Almost all collections prior to the 1960s lacked georeferences provided by the collector. The lack of importance attributed to precise spatial data by the collectors, dating back to the earliest collections made by explorers, is difficult to comprehend since many of these specimens were gathered during geographical surveying expeditions and geological exploration. During the 1960s–1980s, the number of botanical collections in Australia increased rapidly. However, the number of collections that were fully georeferenced remained extremely low (Fig. 1), with the majority of georeferences (at least 90%) calculated by curatorial staff or data processors. Since the late 1980s the increased use of detailed topographical maps, followed by hand-held global positioning systems (since the 1990s), has resulted in a significant increase (75% in 1989–1990) in the number of collections that are fully georeferenced. Although the sample size for the period 1991–2000 is too small to provide an accurate comparison with previous years, there is a noticeable increase in the number of collections fully georeferenced (approximately 90%). The available georeferencing tools mean that the qualitative spatial component of the collections can be readily quantified as part of the collection or curation process. Prior to this, there was no reason for providing georeferences because spatial software was not widely available in herbaria for mapping the distribution of these collections. Up until this period, species distribution maps were usually prepared manually. Until recently, general collectors were more interested in the identity of the collection rather than how it fitted into the known distribution of the taxon. Furthermore, herbarium collection databases were not and are still usually not designed to analyse data. The urgent requirement for data and/or the associated resources required to gather these data, on which environmental management decisions are made, frequently precludes the option of re-surveying an area or distribution of a taxon. It is important to make better use of existing collections and to add a higher level of spatial information into these existing historical collections so that they can be used in applications that require georeferenced data.

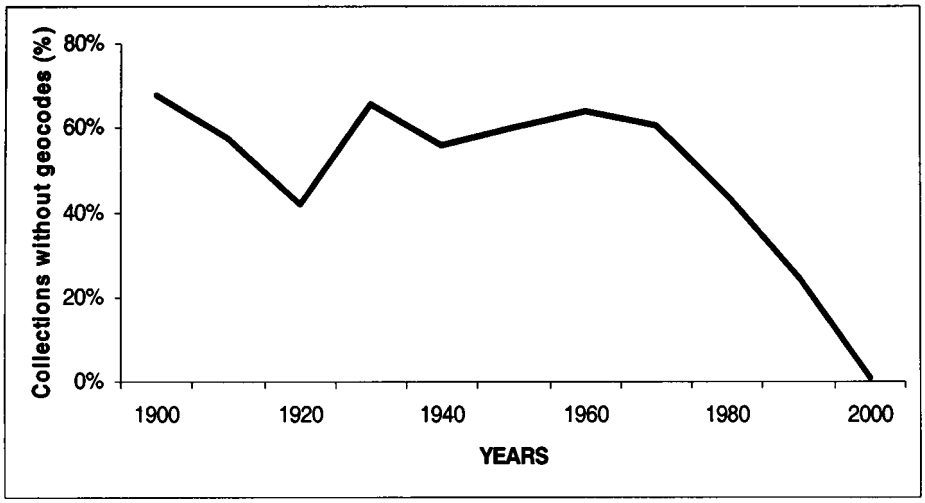

Fig. 1. Percentage of records lacking georeference data taken from a sample of 16 300 herbarium records data processed for the *Australia's Virtual Herbarium* project in the *Rapid* [entry] database (NSW).

It should be noted that given the wide variety and standard of information provided by collectors, the information that is most consistently provided includes the original location of the collection. Other label information is more difficult to interpret across the entire collection because of incompleteness and/or inconsistencies. This information enables the users to answer two important questions, namely, 'What taxa occur in a particular area?' and 'Where does a particular taxon occur?' Both of these questions are fundamental to conservation planning and environmental decision-making.

Impediments to rapid georeferencing

The advent of geographic information systems (GIS) and web-based mapping applications (such as, used by the *Australia's Virtual Herbarium* project) have placed ever-increasing demands on the biodiversity informatics infrastructure to complete digital biological collection catalogues. For example, there are data for approximately 2.3 million specimens available through *The Species Analyst* (Anon. 1998) and *Lifemapper* (Anon. 2002), but less than 30% of these records contains georeferenced locality data. Additional estimates are provided in Table 1. This omission severely limits the value of the specimen data for spatial analyses. Even though most digital data providers recognize the importance of georeferenced specimens, there exist a variety of reasons why these data are not more commonly available. Many of these reasons relate to the cost of this time-consuming task, including the lack of human resources and material resources.

Table 1. Estimates of biological collection size and the percentages digitally catalogued and georeferenced. Note: *The Species Analyst* and *Lifemapper* are both web-based distributed database systems. *NSW Collections* is the botanical specimen database of the National Herbarium of New South Wales (NSW).

Specimen origin	Specimen location	Taxonomic domain	No. of specimens (estimate)	No. digitally catalogued (%)	Number georeferenced (%)
Worldwide	Worldwide	Natural history collections	2.5 billion	< 5	< 5
Worldwide	USA	Natural history collections	0.5–1 billion	< 5	< 2
Worldwide	*The Species Analyst & Lifemapper*	Natural history collections	2.3 million	100	33
Worldwide	USA	Amphibians & reptiles	4.3 million	90	13
Australia	Australia	Vascular plants	6.5 million	22	14
New South Wales	*NSW Collections*	Vascular plants	400000	50	30

At NSW, the process of digitally cataloguing specimens has been ongoing since 1985, with data entry typically requiring 5–30 minutes per specimen. More than half of this time is usually spent on manual georeferencing using paper maps or electronic gazetteers. More recently, desktop geographic information system (GIS) software is being introduced. By using an innovative collaborative georeferencing environment and a well developed set of georeferencing guidelines (Wieczorek 2001), participants in the 'Mammal Networked Information System' (MaNIS 2001) have achieved efficiencies and economies of scale resulting in georeferencing rates of 20 localities per hour. Since MaNIS participants are georeferencing unique localities, each of which refers to an average of about five specimens, the actual rate of specimen georeferencing is closer to 100 per hour. Preliminary results of the prototype automated georeferencing tool (as described here) with MaNIS localities and methodology suggest that the MaNIS georeferencing rate could be increased by an order of magnitude.

Automated georeferencing prototype

Locality information in biological data sets is by no means standardized, but it is to some extent similar across collections, making the task of automated parsing tractable. Similarities notwithstanding, there remain a number of interesting challenges in parsing biological collection locality information. A few examples of textual localities illustrating these challenges are shown in Table 2.

Table 2. Sample textual localities and the challenges posed.

Example textual locality	Challenge posed
Wakarusa, 24 mi WSW of Lawrence	Two or more locations descriptors that are not exactly the same place
Moccasin Creek on Hog Island	Topological nesting
Bupo [?Buso] River, 15 miles [24 km] E of Lae	Complex interpretative description
16 km (by road) N of Murtoa	Linear feature measurement
On the road between Sydney and Bathurst	Linear ambiguity
Southeast Michigan	Vague localities
Yugoslavia	Political borders change over time
British North Borneo	Historical place names

Users of biological collection data have differing needs for spatial resolution, implying that accuracy and precision information need to be captured as integral components of the georeferencing process (Wieczorek 2001). In addition, this information will be essential if biological collection data are to be integrated with other scalable spatial data.

The development of an operational georeferencing prototype using PERL: the Practical Extraction and Reporting Language (Christiansen 2001) to batch process locality data from biological collections is nearly complete. In the original version of this prototype, a query by taxon (e.g., species) to the National Herbarium of New South Wales specimen database (*NSW Collections*) was processed and the results were returned either as a summary report of map-coordinates in HTML, or through a web-based mapping interface. Figure 2 illustrates a sample result in the mapping interface based on a query to the *NSW Collections* database on the genus *Elatostema* (Urticaceae – the stinging nettle family) from Papua New Guinea. Six points (in black) were mapped from data records in which the latitude and longitude were stored in the database. Points in white represent locality data georeferenced by automatic analysis of the descriptive locality data only. In this example, one locality was manually incorrectly georeferenced (as 16°55′S; 155°56′E) (refer black point in the Coral Sea, directly south of Bougainville Island – approximately 1 000 km south). This collection was actually from Buin, Bougainville Island (06°44′S; 155°56′E). Therefore, the manually derived latitude had been either incorrectly calculated or incorrectly data processed. In this example, the automated georeferencing protocols were able to provide some simple error-checking for collections from similar localities. However, it is important to realise that the automated georeferencing protocols are neither more nor less accurate than manually derived georeferences. The main advantage for automating georeferencing is that georeferences can be generated much more rapidly than by manual techniques.

The Western Australian Herbarium (PERTH) have implemented a hybrid system that includes a manual parsing of the descriptive locality into its components during data entry, followed by an automated conversion of these components by appropriate algorithms (P. Goia, pers. comm., 29 November 2002). This approach is expected to remain a useful strategy because manually parsing the locality statement then automating the calculation, is substantially faster than manually calculating the georeference. However, it relies on a high level of operator spatial knowledge.

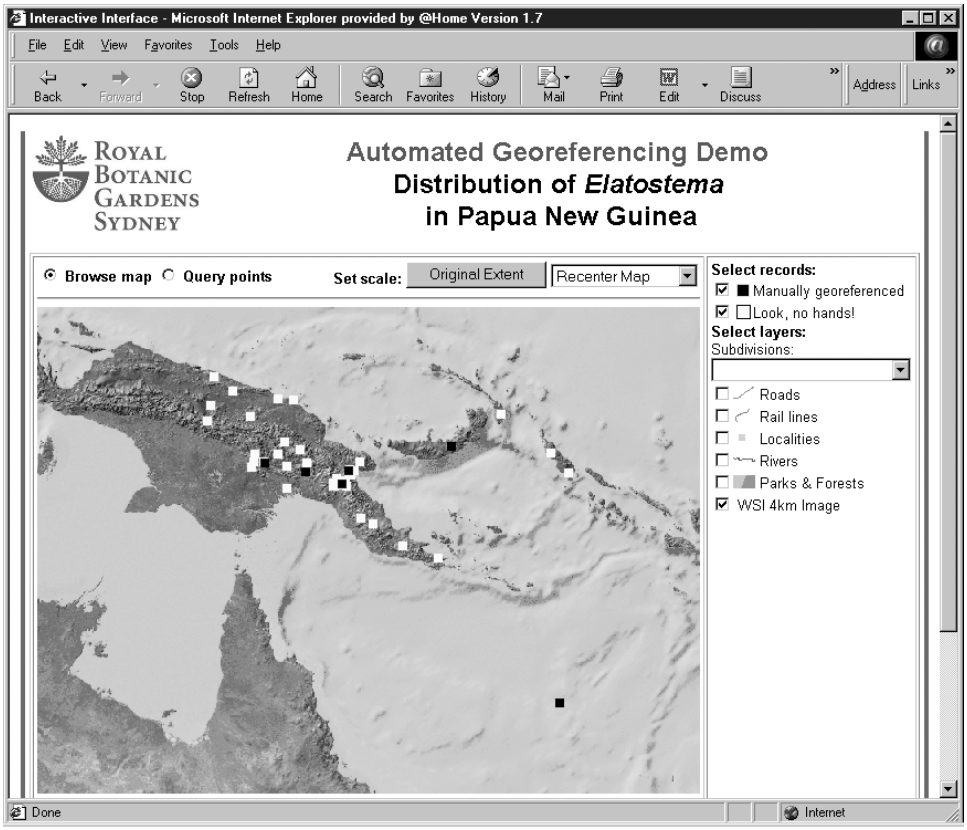

Fig. 2. Web-based mapping results from a specimen database query. Six records (black squares) have mapping coordinates stored in the database, one of which is obviously in error. Automated georeferencing results are shown in white.

The original prototype was based on specimen georeferencing of plants from the Malesian region. However, intensive collaborative interest in georeferencing tools has recently widened the geographic scope of the project to include Australia, through the NSW node of the AVH project. The MaNIS project has recently also started testing the prototype to pre-process mammal biological collections. A georeferencing service has been made available to the MaNIS project through a CGI portal (Beaman 2002) that accepts biological collection data in delimited text format and returns georeferencing results in tab-delimited format.

The automated georeferencing process

The multi-step process involves:

- Pre-processing text for language, locale or project specific anomalies (eg. standardising abbreviations).
- Phrase analysis – the description is compartmentalized by punctuation, prepositions, and stop words into separate phrases. Each phrase is analysed independently.

- Text parsing and pattern matching using regular expressions will involve detecting feature types (e.g., National Park, Island), place names, and their inter-relationships. The web-based prototype (Beaman 2002) queries USA places names from a Geographic Names Information System (USGS 1997). These data are stored locally in a PostgreSQL database.
- Calculation of geographic offsets (e.g., 2.5 km WNW of …).
- Returning results. For the MaNIS project this is done as delimited text. In addition, a PERL/CGI/SOAP interface is currently being designed (by one of us – RSB) for interoperability with *Lifemapper*.

Results of initial tests for MaNIS data from San Bernardino and Placer counties in California (USA), representing the collective data for 17 institutions, are summarised in Table 3 (below). The automated georeferencing results from two samples data sets queried from *The Species Analyst*. A preliminary comparison of the automatically georeferenced Placer County data with georeferences determined manually shows that the mean deviation between results is 0.037 degrees, or about 5.4 km at the latitude of the County. Multiple place-name matches may indicate a higher level of ambiguity in the original text expression. Automated georeferencing of terrestrial vertebrates and birds accurately georeferenced 77–87% of the collections sampled (Table 3), with only 12–26% unable to resolve the descriptive locality because more than one possible georeference was possible (Table 3). These initial tests show that the prototype specimen georeferencing web service is quite promising, but further development and analysis of the results are required to improve the accuracy of the protocol.

Table 3. Summary statistics from automated georeferencing sample data.

Collection(s)	Taxonomic Group	Records tested	Records matched	Records with > than one match
MaNIS	Terrestrial vertebrates	4973	3829 (77%)	861 (17%)
Cornell	Birds	10000	8125 (81%)	1009 (12%)
Univ. of Michigan	Birds	7861	6917 (87%)	2029 (26%)

Conclusion

Individual institutions housing biological collections typically lack the resources or informatics expertise to meet the challenges of georeferencing alone. A community-wide georeferencing solution, equally accessible to all collection-holders, will provide cost-effective added value through economy of scale. Given the scope of both the specimen locality data and the demand for it in a readily usable form, efficiency and accuracy are of prime importance in the task of georeferencing. A multilingual, automated solution that is accessible to individual data providers as well as being interoperable with existing data networks and digital gazetteer services will offer the greatest possible benefit. The solution will need to encompass natural language processing (geo-parsing) to interpret descriptive localities, place-name lookup to register localities with known geographic coordinates, error analysis to self-document uncertainties in the resulting geographic descriptions, and data validation tools with which to analyse the results of georeferencing determinations.

The methodology has shown initial promise for georeferencing biological specimen data, as indicated in Table 3. Based on success rates of up to 87%, future enhancements will initially build upon this model. Research on self-learning for the text parsing module is planned, providing for greater extensibility into other languages. Development will be tailored to the biological collections community through web services made interoperable with other biodiversity community analysis and visualisation applications. Most significantly, error analysis and validation need to be included as integral components of these services. Furthermore, error analysis and validation provide extensive opportunity for testing and further refinement of the geoparsing engine.

The ultimate goal of the project (via the *BioGeoMancer* project, Beaman 2002) is to provide georeferences for natural history collections in a biological context. The research and development for the *BioGeoMancer* project will encompass several research areas within the digital libraries framework. These areas include

- Natural language processing
- Geospatial integration
- Spatial error analysis
- Linkage between geospatial base data and biological classification
- Interoperability between organizations with discordant data standards
- Scalability

In order to achieve our goal, we have assembled a collaborative partnership that provides links to several disciplines: botany, zoology, biodiversity informatics, and computer science engineering. Although the *BioGeoMancer* partners have developed expertise in many disciplines independently, this project provides a unique opportunity to combine resources in a cohesive, productive manner.

Acknowledgments

We thank Gary Chapple and Ken Hill (both NSW) for their generous support and considerable technical advice and assistance during the initial development of this project in Australia. Paul Goia (Department of Conservation and Land Management, Western Australia) kindly reviewed this paper for us. One of us (R.S.B. – as principal investigator) was supported by a three-year National Science Foundation grant (DBI –9974217) at the Royal Botanic Gardens Sydney, Australia (NSW) and the University of Kansas, Lawrence, USA (KUNHM-BRC). This generous support is gratefully acknowledged.

References

Anonymous (1998) *The Species Analyst* [http://apps.internet2.edu/sept98/species.htm] and [http://habanero.nhm.ukans.edu/].

Anonymous (2001) *Council of Heads of Australian Herbaria (CHAH)* [http://www.anbg.gov.au/chah/chah-intro.html].

Anonymous (2002) *Lifemapper* [http://www.lifemapper.org].

Barker, W.R. (1998) *The Virtual Australian Herbarium: a cooperative flora information system being developed by Australian herbaria* [http://plantnet.rbgsyd.gov.au/HISCOM].

Beaman, R.S. (2002) BioGeoMancer [http://biogeomancer.org].

Christiansen, T. (2001) *O'Reilly perl.com – The source for PERL* [http://www.perl.com/]

Conn, B.J. (1996) *HISPID3 – Herbarium Information Standards and Protocols for Interchange of Data*, version 3 (Royal Botanic Gardens Sydney) [http://plantnet.rbgsyd.gov.au/HISCOM].

Conn, B.J. (2000) *HISPID4 – Herbarium Information Standards and Protocols for Interchange of Data,* version 4 (Royal Botanic Gardens Sydney) [http://plantnet.rbgsyd.gov.au/HISCOM].

Conn, B.J. & Brooks, A.K. (1998) *HISCOM – Herbarium Information Systems Committee* [http://www.rbgsyd.gov.au/HISCOM]

Duckworth, W.D., Genoways, H.H. and Rose, C.L. (1993) *Preserving Natural Science Collections: Chronicle of our Environmental Heritage* (National Institute for the Conservation of Cultural Property: Washington, D.C.).

GBIF (2002) *The Global Biodiversity Information Facility* [http://www.gbif.org/index.html].

Maier, D., Landis, E., Cushing, J., Frondorf, A., Silberschatz, A., Frame, M. & Schnase, J.L. (eds) (2000) *Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystem Informatics* (NASA Goddard Space Flight Center). [http://bio.gsfc.nasa.gov.].

MaNIS (2001) *The Mammal Networked Information System* [http://dlp.cs.berkeley.edu/manis].

USGS (1997) *Geographic Names Information System* [http://nsdi.usgs.gov./products/gnis.html].

Wieczorek, J.R. (2001) *MaNIS: Georeferencing Guidelines* [http://dlp.cs.berkeley.edu/manis/GeorefGuide.html].

Wilson, E.O. (2000) *On the future of conservation biology*. Conservation Biology 14(1): 1–3.