ARTICLES

D-Lib Magazine
May 2004

Volume 10 Number 5

ISSN 1082-9873

Determining Space from Place for Natural History Collections

In a Distributed Digital Library Environment

Reed Beaman
Peabody Museum, Yale University
<reed.beaman@yale.edu>

John Wieczorek
Museum of Vertebrate Zoology, University of California, Berkeley
<tuco@socrates.berkeley.edu>

Stan Blum
California Academy of Science
<sblum@calacademy.org>

Abstract

More than a billion biological specimens have been collected, preserved, and deposited in the permanent collections of museums and herbaria around the world. These specimens are the foundation of our knowledge about biological diversity, past and present. Researchers in biodiversity informatics are engaged in providing digital access to the basic biodiversity data associated with specimens, as well as new software tools and services that will create novel research opportunities for ecological analysis, predictive modeling, and synthesis. Greater access to structured biodiversity information also directly benefits applied areas of conservation and resource management. Digitizing the data associated with a billion specimens is an enormous task, and much of it still lies before us. Already, however, tens of millions of specimen records have been captured in collection management systems that represent a solid foundation for comprehensive digital libraries in the museum community.

Of the various classes of information linked to biological specimens, geospatial coordinates used for mapping species distributions are among the most widely demanded by the scientific community and the general public. Providing these coordinates (geospatial referencing) has proven a significant challenge. Nearly every specimen is associated with a georeference, most often a textual representation of the place from whence it came, but few carry with them quantitative geospatial coordinates. We address here advances made in the task of geospatial referencing for biological collections.

Background: the case for collections

Managing finite natural resources is among the greatest challenges in this century facing biologists, information scientists, managers, economists, and policy-makers. Meeting this challenge is critical for sustaining human growth and prosperity, maintaining economic stability, and improving the quality of life for all species. Knowledge of biological diversity is fundamental to natural resource management, and the urgency for this knowledge ever increases as we convert the final tracts of the natural landscape into human-managed systems. Much of this knowledge remains locked in physical archives and on library shelves. In order for biodiversity management to keep pace with the rate of resource exploitation, digital access to biological diversity information is imperative in this decade.

Specimens, the primary objects in a biological collection, are of principal importance in documenting the spectacular biological diversity of plants and animals on this planet. Specimen data document the identities, habitats, histories, and spatial distributions of the roughly 1.75 million described species of life on earth, and provide the fundamental resource for identifying the estimated tens of millions of species that remain to be discovered and described (Wilson, 2000). More than 250 years of natural history exploration have contributed to a vast worldwide collection of over one billion biological specimens. These specimens are essential pieces of the puzzle for answering four very basic questions: How many species are there? What are their names? How are they related to each other? Where are they found? Addressing these questions has been the domain of biological systematics (or taxonomy) since the time of Linnaeus (ca. 1735). Yet, to this day, answers to these four questions are largely incomplete. Biodiversity informatics facilitates the effort to resolve these questions, seeks to create new efficiencies for data capture and retrieval, and provides novel approaches for interdisciplinary analysis and visualization.

The focal topic of this article is our access to digital data, particularly geospatial data, associated with biological collections. The digital acquisition, integration and application of biological collections data are increasingly viewed as fundamental to biodiversity research, education, and natural resource management. Many specimens linger for decades on the shelves of thousands of individual collections before a specialist gets an opportunity to study them. This deficit exists in large part because access to collections is limited by their distributed nature.

The ABCs of accessing biological collections

Providing digital access to biological collections is enabled by two critical technologies: (1) the digital capture of natural history collection data, and (2) the ongoing development of a cyberinfrastructure that links these digital collections into a global distributed digital library environment. A great deal of progress has been made recently on developing the infrastructure to provide digital access to data. Models for community solutions for distributed data access are evident in numerous projects such as the Global Biodiversity Information Facility (GBIF, 2002), Australia's Virtual Herbarium (AVH, 2002), the Biological Collection Access Service for Europe (BioCASE, 2003), the Comisi ón Nacional para Conocimiento y Uso de la Biodiversidad (CONABIO, 2002), the European Natural History Specimen Information Network (ENHSIN, 2000), FishNet (2000), HerpNET (2003), and the Mammal Networked Information System (MaNIS, 2001), which are establishing distributed Internet-based information systems for accessing, integrating, and conducting predictive biotic modeling on specimen-based biodiversity data. Significant technological development, funding, and effort have gone into these projects, primarily for distributed information retrieval, data capture, and human-mediated geospatial referencing.

Space is the place

Every specimen has an associated core set of data that answers what, when, how, and where. A taxonomic

name indicates the object's biological identity, the "collecting method" specifies how it was collected, and a list of one or more "collectors" indicates who collected it and when. While modern collectors are usually able to provide geospatial references in the form of coordinates derived from the Global Positioning System (GPS), the vast majority of specimens collected more than 10 years ago have only a natural language description, a textual georeference, as their specification of place.

Quantitative geospatial references are critical elements of biological collection data, because they enable us to integrate information within and beyond the biological domain. Geospatial referencing provides the means to link specimen data to the rapidly growing body of spatial environmental data for interdisciplinary research into complex phenomena. For example, access to geospatially referenced data from specimens provides a quantitative basis for biodiversity analyses and predictive niche modeling and for determining sampling densities of various sites (e.g., Figure 1):

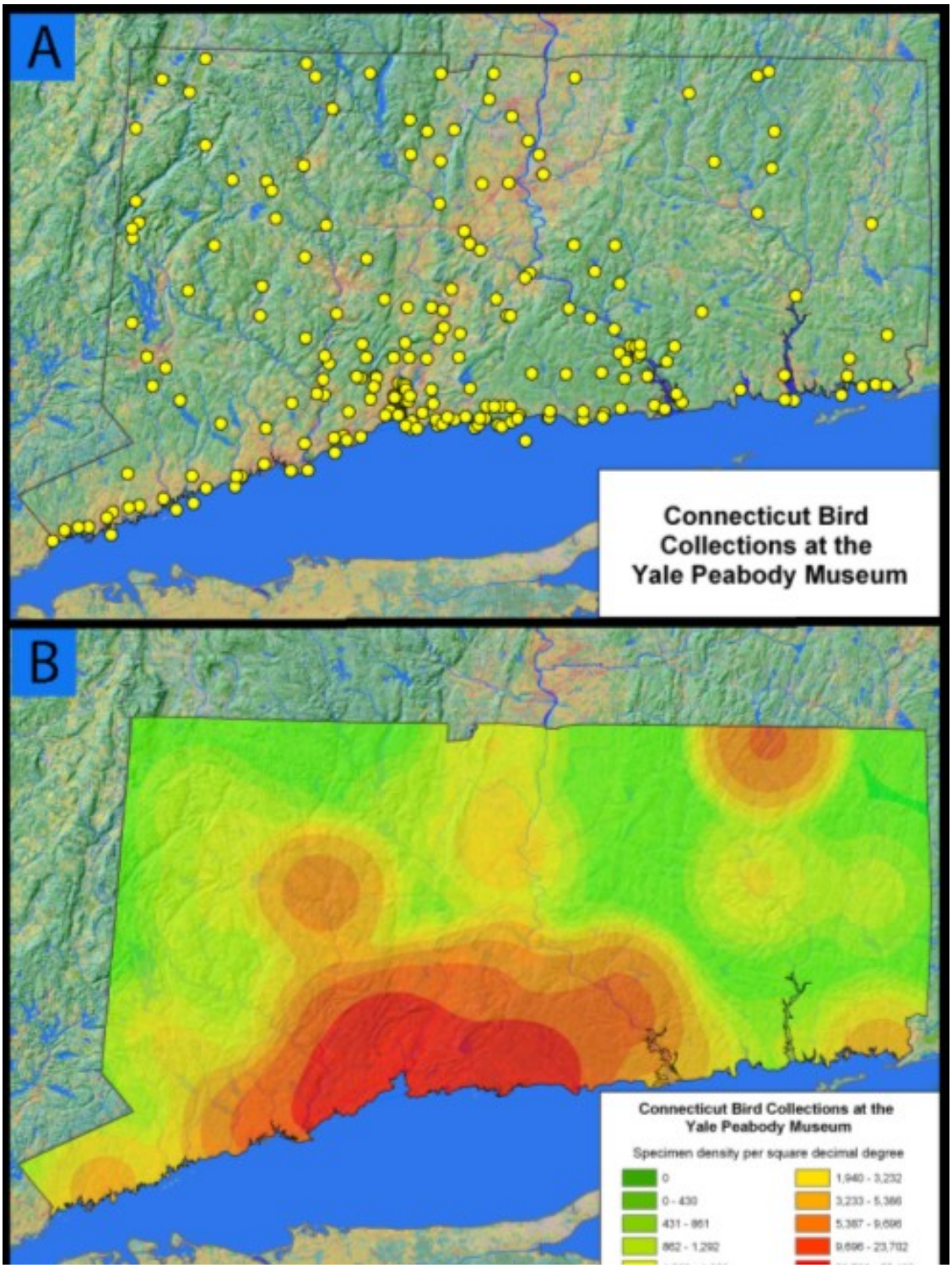

Figure 1. Comparison of localities sampled (A) with density of specimens sampled (B). Density plots provide a more informative visualization of sampling completeness. Specimen data provided by Kristof Zyskowski, Peabody Museum.

Because most collecting localities are in textual form, the museum community is facing an enormous task of converting natural language georeferences of place to geospatial references of space. The task of retrospectively deriving space from place, whether carried out using paper maps and rulers or with desktop GIS tools, is non-trivial when considered for all collections worldwide. This can represent an additional 25-100% overhead on the cost of digital data capture. A challenge therefore arises about what we can do to accelerate this process.

Locality information in biological datasets is by no means standardized, but it is to some extent similar across collections, making the task of parsing tractable. Similarities notwithstanding, there remain a number of interesting challenges. A few examples of textual locality statements illustrating these challenges are shown in Table 1.

Example textual georeference	**Challenge posed**
Wakarusa, 24 mi WSW of Lawrence	Two or more locations described 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 grammar
16 km (by road) N of Lawrence	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 placename

Table 1. Sample textual georeference and the challenges posed.

We summarize below a semi-automated methodology based on collaboration and web services that are accessible to individual data providers and are interoperable with existing data networks. The solution encompasses natural language processing to interpret descriptive localities (geoparsing), placename matching to digital gazetteer data 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 visualize the results of geospatial references. This solution provides cost-effective added value through an 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 our task of geospatial referencing.

Collaborative geospatial referencing

Individual institutions housing biological collections typically lack the resources or informatics expertise to meet the challenges of georeferencing alone. By designing an innovative collaborative geospatial referencing methodology, participants in the National Science Foundation funded Mammal Networked Information System (MaNIS, 2001) have achieved mean geospatial referencing rates of 16 localities per hour. Since MaNIS participants are referencing localities applicable to an average of about five specimens, the actual rate of specimen geospatial referencing is closer to 80 per hour (Wieczorek *et al.*, 2004).

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. In addition, this information will be essential if biological collection data is to be integrated with other scalable spatial data (Withey *et al*., 2002). One of the key aspects of collaborative geospatial referencing for MaNIS is to provide an uncertainty measured as a maximum error radius around a geospatial reference point using a specialized *Georeferencing Calculator* (Wieczorek, 2001b) implemented as a Java applet. Uncertainty values are calculated based on the characteristics of spatial relationships described in the natural language georeference, geospatial extent of a place, knowledge (or lack of knowledge) about geodetic datums, and levels of implied precision. The algorithms embedded in the *Georeferencing Calculator* are documented in the MaNIS Georeferencing Guidelines (Wieczorek, 2001a).

MaNIS data are currently available through a distributed network of providers accessible through one or more presentation layers or portals. This distributed network is implemented through a schema independent protocol, Distributed Generic Information Retrieval (DiGIR, 2001). DiGIR was developed in part through the effort of co-authors Wieczorek and Blum as an essential means of sharing collection data in the biological community.

Web services

Digital data capture rates can and must be improved. While there are now tens of millions of digital data biological collection records available through the various sources mentioned above, it is a small percentage of the total. Development of automated and semi-automated geospatial referencing will greatly enhance the efficiency and efficacy of geographic data capture. BioGeomancer (Beaman, 2003) is a conversion engine for automating the conversion of references of place into geospatial references for natural history collections. The intent is to be able to handle language constructs such as those listed in Table 1. In its current incarnation, this BioGeomancer can parse English language placename descriptions and provide a set of latitude/longitude coordinates associated with that description. It provides offset calculations for when a collection is textually georeferenced at a given distance and cardinal direction from a named place.

The basic steps in the process include pre-processing text for language, locale or project specific anomalies (e. g., standardizing abbreviations) followed by one or more iterations of a phrase analysis algorithm. The description is segmented by punctuation, prepositions, and stop words into separate phrases and each phrase is analyzed independently. Within each of the phrase analysis iterations, text parsing and pattern matching using regular expressions involves detecting feature types (e.g., National Park, Island), placenames, and their inter-relationships.

Candidate placenames are passed as a query to a local PostgreSQL database of gazetteer entries from the Geographic Names Information System (GNIS) (USGS, 1997) for U.S. data and National Geospatial-Intelligence Agency GEOnet Names Server (NGA, 2003) for all other countries. Geographic offsets (e.g., 2.5 km WNW of...) are detected and calculated relative to returned place names. The query interfaces allow either single or batch processing of multiple locality records. Web services are implemented through a Simple

Object Access Protocol (SOAP, 2000) interface for interoperability with other applications. Cross-platform interoperability has been tested with client applications written in various languages including Perl, PHP, Java, and .NET. For human-based access, a simple web form mediates as a wrapper for SOAP requests, post-processing responses with XSL stylesheets into HTML or passing the responses to an interactive web mapping interface.

A response summary and matched record detail are returned. Ambiguities in the response are caused by certain conditions, for example, the original designation of place includes more than one placename (e.g., "Micanopy, 11 miles S of Gainesville") and placename lookup in the gazetteers results in multiple places with the same name (the *homonym problem*). These ambiguities are handled by including an uncertainty estimate in the summary. This is calculated by placing a bounding box around all returned points, then measuring the distance from the centroid to a corner of the box. In a semi-automated environment, the error estimate can be used to automatically flag geospatial references that exceed a threshold for ambiguity.

Giving back: event gazetteers

Resolution of homonyms and synonyms in gazetteers is an area where knowledge derived from biological collections may contribute to gazetteer knowledge bases. For example, Madagascar presents an unusual homonymic challenge for geospatial referencing. There are 52 matched records for the place "Manakana" in Madagascar in the GeoNet Name Server. A bounding box placed around these matches has a distance of 1,368 km from corner to corner. Numerous synonymic variations in spelling also compound the ambiguity inherent in retrospective geospatial referencing. Further, numerous rural communities have moved in recent historical times and have retained the same name.

Thus, it will be important to describe these data beyond placenames and geographical feature types. Traditional gazetteer schemas generally define a placename, a feature type, coordinates, and some indication into which larger administrative entities a place fits. The Alexandria Digital Library (ADL) Gazetteer includes concepts of time-stamps (to address the issues of name, boundary, type, and relationship changes over time) and further descriptive information. We find that it will be extremely informative to extend specialized gazetteer schemas to include the concept of *collection events* (the name of a collector or expedition and a date of collection). As an example of how this might be useful, historical specimen records that we have checked often do not include higher order administrative units (e.g., states, provinces, counties, districts, etc.) but do include the collector's name and date of collection. By linking the "who, when, and where" of collecting activities, the development of *collection event gazetteers* can solve many of the problems of spatial homonymy. Homonyms can be distinguished by knowing the collectors' identities and location during a particular time period. These are data elements that are not only found on specimen label data, but also with notes, logs, diaries, and maps associated with collections.

Systematic transformations

Biological specimens are the primary source of information for measurements and observations used to infer relationships in the tree of life. Knowledge is also built on diverse additional resources in natural history collections, including illustrations, photographs, collectors' original notes, diaries, and maps. In the same way that medical informatics is able to provide distance-based diagnosis and treatment, access to biological collection data creates a virtual *collaboratory* for systematic study where information synthesis and assimilation that used to take days or weeks can occur in a few seconds. Providing this access to the scientific community facilitates systematic research and novel types of biogeographic analysis through GIS-based mapping, analysis, and visualization of species distributions. Indeed, these changes reach well beyond the systematics community. Until very recently, only specialists in systematics had access to this treasure trove of biological data. Improving access to biological collections in related disciplines adds tremendous value to the

collections and our ability to synthesize biological data into associated domains. For example, scientists with improved access to species distribution data can improve the accuracy and precision of predictive distribution modeling for invasive species, pests, and disease-related research. Conservation and resource managers can better target areas of high biodiversity for protection and management. Students and educators can have access to the same information available to researchers scaled through appropriate views.

The biodiversity community is by necessity embracing new technologies that transform how systematists conduct their science and how biological collections are managed. For example, an enormous transformation in systematic knowledge has occurred over the last few decades through the use of molecular (e.g., DNA sequencing) and computationally intensive methods of phylogenetic interpretation. This transformation has primarily affected our understanding of organismic relationships, amplifying our ability to count, classify, and name organisms, and has transpired in the laboratory at the molecular and algorithmic levels. Molecular systematics has greatly modified the scientific landscape for understanding our biological heritage. Still, there is the critical need to link molecular data to real organisms. Specimens serve as the vouchers for molecular biology, ecology, physiology, and pathology, representing the crucial comparative means for linking names to organisms. The practice of biology without vouchers creates an undocumented link between experimental organisms and scientific names. The names, distributions, and economic values are often the only aspects of most species that the lay community (including legislative bodies) will ever see.

Conclusions

Geospatial referencing is one of the new technologies that promises to unlock the biological and geographical knowledge contained in the world's natural history collections. It adds tremendous value to the use of biological specimens in the biological community and beyond, by allowing us to integrate this rich source of biological knowledge with geospatially referenced data in other domains through digital library architectures. Significant challenges remain ahead for improving technologies and methodologies that spatially enable natural history collections. As a foundation of knowledge for our planetary biological diversity, we must meet these challenges as a priority, and geospatial referencing is a prerequisite for the scientific synthesis essential for understanding and managing the complexities of our natural environment.

References

AVH. 2002. Australia's Virtual Herbarium, (http://www.anbg.gov.au/avh.html).

Beaman, R. 2003. Biogeomancer: Automated georeferencing for natural history collections, (http://www.biogeomancer.org).

BioCASE. 2003. Biological Collection Access Service for Europe, (http://www.biocase.org).

CONABIO. 2002. Comisión Nacional para Conocimiento y Uso de la Biodiversidad, (http://www.conabio.gob.mx/).

DiGIR. 2001. Distributed Generic Information Retrieval, (http://digir.sourceforge.net).

ENHSIN. 2000. European Natural History Specimen Information Network, (http://www.nhm.ac.uk/science/rco/enhsin/).

FishNet. 2000. The FishNet Distributed Biodiversity Information System, (http://habanero.nhm.ukans.edu/

fishnet/).

GBIF. 2002. The Global Biodiversity Information Facility, (http://www.gbif.org).

HerpNET. 2003, (http://www.herpnet.org).

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

NGA. 2003. National Geospatial-Intelligence Agency GEONet Names Server, (http://earth-info.nga.mil/gns/html/index.html).

SOAP. 2000. Simple Object Access Protocol (SOAP) 1.1, (http://www.w3.org/TR/SOAP/).

USGS. 1997. Geographic Names Information System (GNIS), (http://gnis.usgs.gov/).

Wieczorek, J.R. 2001a. MaNIS: Georeferencing Guidelines, (http://elib.cs.berkeley.edu/manis/GeorefGuide.html)

Wieczorek, J.R. 2001b. Georeferencing Calculator, (http://elib.cs.berkeley.edu/manis/gc.html).

Wieczorek, J., Guo, Q., and Hijmans, R. In review, 2004. The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*.

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

Withey, A., W. Michener, and P. Tooby, (eds). 2002. *Scalable Information Networks for the Environment (SINE): Report of an NSF-sponsored workshop*, San Diego Supercomputer Center, October 29-31, 2001. 65 pp.

Top | Contents
Search | Author Index | Title Index | Back Issues
Previous Article | In Brief
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/may2004-beaman