NSF Georeferencing U.S. Fish Collections

Results of Prior Support (Lead Institution)

BDI–0417001: "Collaborative Research: Building the Information Community Infrastructure - A Test Case Implementation for Ichthyological Collections." $290,375, Aug. 2004 to Jul. 2009. This five-year, continuing award was funded in collaboration with researchers at the University of Kansas. The aim of the project was to build an information community infrastructure for natural history collections using distributed fish collection databases as a test case. Tulane's role in the project was to design a collaborative georeferencing system that harvested locality data from collections using the DiGIR 2 protocol, georeferenced the locality data using GEOLocate, and performed similarity matching of collections locations to make the verification process more accurate and efficient. We created a georeferencing server to retrieve locality data from DiGIR providers. We then modified GEOLocate to retrieve location information from the georeferencing server for verification, correction and error assignment, and then return data to the georeferencing server. We also developed a web portal for participant management and query, and mapping of collection data.

BDI–0516312: "GEOLocate World: An Expanded Tool for Georeferencing Natural History Collections." $262,571, Aug. 2005 to Jul. 2008. This project allowed us to expand the geographic scope and services of GEOLocate to worldwide. Other goals of this project included development of Taxonomic Footprint Web Services (incorporated into GEOLocate), and multilingual locality string georeferencing. We have expanded the global coverage of GEOLocate to 234 countries and released an additional high resolution dataset for Australia. We worked with GBIF Spain to improve the quality of data available for Spain and produce language libraries which define locality patterns in Spanish, Galician, Catalan & Basque. These libraries are serving as a model for anyone wishing to customize GEOLocate to other languages. The Taxonomic Footprint Web Services have also been completed and are undergoing testing to evaluate the utility of Taxonomic Footprints and to determine ideal default parameters for building the footprints. This work was presented at the 2006 TDWG meeting in St. Louis, Missouri. We are currently focusing on the impact of synonymies on building taxonomic footprints and how best to deal with this issue.

DBI–0852141: "Improving GEOLocate to Better Serve Biodiversity Informatics" $1,134058, Aug. 2009 to Jul. 2012. This ongoing project is allowing us move from the traditional standalone desktop application to a lightweight desktop application functioning via remote distributed services. Specifically we are developing webservices for natural language processing and georeferencing of collecting event data while improving core algorithms for greater efficiency and accuracy, implementing uncertainty calculations and improving data verification via customizable integration of modern online map services. In addition, client interface components will integrate with the Specify 6 workbench (http://www.specifysoftware.org) to provide accurate georeferencing and verification from within Specify. Specify is a desktop application and database for managing natural history collection data.

Introduction

It is estimated that the number of biological specimens in US museums and herbaria exceeds 500 million (Kristalka and Humphrey 2000). The number worldwide is estimated as approaching 3 billion. These collections serve a critical role in documenting the biodiversity of our plant over space and time. Investments in digitization and networking over the past decade have greatly expanded the role natural history collections play in scientific research. In addition to traditional uses in taxonomic and systematic research, natural history collections are now serving as a resource for research in ecology, conservation and environmental biology. In these studies, the specimens and their associated data constitute an historical record, useful for assessing population expansion or decline, or community change in relation to environmental change. As more and more collections enter the digital world, large-scale interdisciplinary studies based on networked museum data become more and more feasible.

Recent years have witnessed the emergence of "*Biodiversity Informatics*", wherein the Internet and computing technology are being used to bring the vast resources of natural history museums into wider use in biodiversity discovery and other areas of scientific inquiry (The Committees of *Systematics* 1202953 2 *Agenda 2000*, 1994, Sugden and Pennisi 2000, Pennisi 2000, Bisby 2000, Edwards et al. 2000, Soberón and Peterson 2004, Suarez and Tsutsui 2004, Tobin 2004, Page et al. 2005, Johnson 2007, Sarkar 2007). Biodiversity informatics has resulted in the development of numerous software applications for data capture, enhancement and analysis. It has also led to the establishment of large data-sharing portals, such as the Global Biodiversity Information Facility (GBIF, http://www.gbif.org), which has provided researchers access to nearly 200 million biotic records from 317 data providers.

Each data record in a biodiversity database or portal represents a specimen occurrence at a specified time and place. The vast majority of these occurrences were recorded prior to development and wide use of GPS technology, and thus do not include geospatial coordinates (latitude and longitude). Instead, the collection locations are recorded in databases as simple strings of text, with varying levels of specificity, with other, higher-order geographic reference fields (e.g., state/province or ocean). Locality descriptions typically describe a collection location as a position along a road, at a distance from a town or other geographic point of reference, or as lying within a particular township, range and section (TRS) area. In order to relate these localities to electronic maps, the locality description must be converted to geospatial coordinates (latitude and longitude). These geospatial coordinates ('georeferences') are critical elements of biological occurrence records, because they enable researchers to map the collection location and relate the accompanying biological data to other kinds of information (e.g., environmental data) for more complex analyses of biogeographic phenomena (Beaman et al., 2004). In the years since the publication of *Teaming with Life*, the report of the President's Committee of Advisers on Science and Technology (PCAST) Panel on Biodiversity and Ecosystems (PCAST 1998), little has changed to diminish the importance of the report's statement, 'that georeferencing is urgently needed to facilitate the use of natural history collection data for the study of status and trends in biodiversity and ecosystems.

Georeferencing allows species occurrences to be visualized on digital maps in relation to other types of digital information (geology, hydrology, climate, other environmental data) using Geographic Information System (GIS) technology. Across a long enough span of collecting effort, the geographic extent of a species' occurrences describe the species' tolerance of environmental conditions, which define its geographic distribution. In fact, implicit in the species' distribution are many other important ecological and environmental relationships. Thus, simply visualizing species occurrences has the potential to teach us much about basic the biology and geography (biogeography) of organisms.

Traditional methods for georeferencing textual descriptions of collection locations, using hardcopy or even digital maps, are tedious and time consuming. Often, multiple hard-copy maps are needed, and the coordinates have to be entered or copied and pasted by hand. Using digital maps in a GIS environment such as Arc View can simplify the process, but the software can be costly and requires some training. Even with digital maps, localities must first be pinpointed, which can be time consuming.

In February of 2002, with NSF funding, the Tulane University Museum of Natural History (TUMNH) began work on a computing solution to this task, using manually-georeferenced data from the TUMNH fish collection as a test bed. The end result was GEOLocate (http://www.museum.tulane.edu/geolocate), a collection of software applications and services that 1) interpret textual descriptions of locality information, 2) translate them into geographic coordinates, 3) project these coordinates onto digital maps, 4) provide a mechanism of verifying and correcting the displayed location, and 6) allow users to assign polygon-shaped error estimates to the determination. GEOLocate currently exists in three formats: a standalone desktop application (primary application), an online web application and a set of web services for developers.

At the core of GEOLocate is a natural language processing and geocoding algorithm which standardizes the locality information in common terms; parses out distances, compass directions, and key geographic identifiers; then performs database lookups and geographic calculations to determine the geographic coordinates. Vital to this process are the datasets used to identify places. Datasets used by GEOLocate include: the United States Geological Survey's Geographic Names Information System, the National Geospatial-Intelligence Agency's GEONet Names Service, the U.S. Army Corps of Engineers Waterway Mile Marker Database, and internally developed bridge crossing and township, range and section (TRS) databases. The bridge crossing database was derived from the U.S. Census Bureau's 2000
1202953 3 Tigerline Data. The TRS database was generated using the TRS2LL program developed by Martin Wefald (http://www.geocities.com/jeremiahobrien/trs2ll.html). In addition to point-based gazetteers, GEOLocate makes use of linear features, such as river networks, to identify locality descriptions along rivers. Partnerships with the Australian Museum and the Spanish node for GBIF have resulted in the addition of high resolution gazetteers for Australia and Spain.

Output coordinates from the geocoding algorithm are then ranked according to the type of information in the locality string and plotted on a map display for user verification, correction and error determination. Making corrections to a georeferenced locality is as simple as dragging the displayed point to a new location on the map and clicking the "correction" button. GEOLocate also gives users the option to create an error polygon to represent uncertainty for more ambiguous locality descriptions.

TUMNH provides GEOLocate free-of-charge to members of the natural history collections community and researchers interested in using natural history collection data. A website is maintained for providing the user community general information about the tool, technical support, contact information and access to the web and desktop applications.

GEOLocate was originally developed for georeferencing collection localities in North America (U.S., Canada and Mexico) but has now been expanded to the entire world and includes a customizable locality typing system in the desktop application to support languages other than English and buit in support for Spanish, Galician, Catalan and Basque. Over 1500 copies of GEOLocate have been distributed to various natural history collections and researchers around the world. Feedback received from users and citations in published research suggest that GEOLocate is a highly accurate tool for georeferencing data from a broad spectrum of taxonomic collections (Murphey et al. 2004, Monfils and Prather 2007, Demastes et al. 2007, Lutrell et al. 2007). Among the past and present users of the software are the American Museum of Natural History Invertebrate Collection, the Southeast Regional Network of Expertise and Collections (a consortium of southeastern U.S. Herbaria), Solanaceae Source, and participants in Herpnet (a network of amphibian and reptile collection databases). GEOLocate web services also allow developers to integrate georeferencing capabilities directly into their applications and databases. Specify, Arctos, Tropicos & Digitarium are among the projects making use of GEOLocate services. During the first half of 2011 GEOLocate webservices averaged 10,400 georeferencing requests per day.

A collaborative grant funded in 2004 allowed the development team to expand GEOLocate's capabilities to include online community-based georeferencing (http://www.museum.tulane.edu/coge/) using Distributed Generic Information Retrieval (DiGIR, http://digir.sourceforge.net/) data providers as data sources, and the FishNet2 community as a test case (http://www.fishnet2.net/). This project funded the development of a GEOLocate-based collaborative georeferencing framework, but did not include support for actual georeferencing. **We propose to maintain and expand the existing FishNet2 network by adding more providers (doubling the number of available records) and using GEOLocate's Collaborative georeferencing platform to assign geographic coordinates to all specimen localities within the network and standardize geographic entities across all data providers. We will also evaluate methods for determining geographic uncertainty in relation to aquatic collection localities.** The proposed work will significantly facilitate and broaden access to the FishNet2 data consumers.