Proposal to the Atherton Seidell Endowment for Digitizing and Disseminating the *Biologia Centrali-Americana*

Smithsonian Institution Libraries
May 6, 2002

Abstract

The Smithsonian Institution Libraries (SIL) requests $594,000 from the Seidell Endowment for the first phase of a multi-phase project that will create an extraordinary new set of resources and knowledge tools in electronic form for biodiversity studies centered on Mexico and Central America. The funds will be used for the first phase, which will cover 2.5 years, to scan, rekey, and code in eXtensible Markup Language (XML) the full text of an important and out-of-print scientific work, *Biologia Centrali-Americana* (BCA). The coded text will be linked to other vital biological datasets including the NMNH collections information system. This project will benefit everyone who deals with biodiversity in Mexico and Central America, including biologists, conservation groups, land planners, natural resource managers, and quarantine officials. The BCA is essentially an early database of biodiversity of Mexico and Central America and thus forms the basis for subsequent advances in knowledge, as well as the foundation for studying changes caused by natural or human factors (e.g., invasive species). The contents are derived from scientific surveys and explorations conducted during the latter part of the 19th century and early 20th century. Many of the leading biologists of the time provided specimens and descriptions for the many volumes. The illustrations are in many cases the only images that exist of the biota of the region. This project will also be the keystone and model for several other major bioinformatics projects to be pursued by the leading biological repositories in the world, which will speed the pace of scientific investigation into the nature of the rapidly changing natural environment.

Background

At an October 2001 conference organized by the Smithsonian Institution and funded by the Andrew W. Mellon Foundation, *Toward Collaborative Biodiversity Informatics: Mobilizing Collections and Research Data*, key representatives from the American Museum of Natural History, The Natural History Museum (London), the Royal Botanic Gardens (Kew, UK), Missouri Botanical Garden, and the Smithsonian Institution (National Museum of Natural History, Smithsonian Tropical Research Institute, and the Smithsonian Institution Libraries) met to address the problems of managing and accessing the information embedded in large biological repositories.

Natural history museums and similar large biological repositories and their libraries hold a wealth of inadequately accessible resources that describe and explain the diversity and depth of life on earth. Mining these data for research, conservation, drug discovery, protected area management, disease control, etc., is difficult, time consuming, and often leads to redundant efforts. What should be a seamless, open "book" of knowledge consists, instead, of disparate, unintegrated sets of data - some in electronic form but

most still on paper, published and unpublished. The conference focused on the four following types of biological datasets:

1. *Specimen collections* in large repositories. Many large biological repositories, such as the Smithsonian's National Museum of Natural History, are converting manual records about these collections into integrated electronic collection information systems. Core data elements typically include taxonomy (especially species name), collection locality, collector, and identification number(s) (catalogue number, collector's number, accession number, etc.). Images, especially of type specimens, are also a priority.

2. *Taxonomic databases* that record the names (including type and literature references that validate the names in terms of the governing International Codes of Nomenclature), classification, synonymy, geographic distributions and relationships of biological organisms.

3. *Published taxonomic literature*, including journal articles, monographs, and other forms of publication that name and describe taxa (according to the Codes of Nomenclature), details of collection, and other information.

4. *Geographical information systems* (GIS) that link geographic place names and other geographic data elements with precise geospatial coordinates. Because names have changed, often repeatedly, in the more than two centuries that such animal and plant specimens have been collected, GIS are important for pinpointing the exact locations where such specimens were found. Once large numbers of specimens have been georeferenced, aggregate studies may be performed, such as species distribution over time.

These datasets are part of a larger, worldwide effort to enable easy access to the complete range of data required to understand individual species and their environmental and evolutionary relationships. This will require the establishment of cross-linkages between and simultaneous access to, data sets from such information sources throughout the world.

To deliver an immediate product, provide a testbed for developing protocols and technologies for linking these datasets in a coherent, usable way, and act as a model for further projects, the conference participants agreed on an immediate high-priority project based on the *Biologica Centrali-Americana* (BCA), one of the most important compendia of the flora and fauna of Mexico and Central America.

A repeated message from those interested in conservation of biodiversity around the world, especially those in biodiversity rich but resource poor countries, is the need for taxonomic information. This is necessary for a wide range of environmental management and conservation purposes, as well as being a basic tool for education and enjoyment of the natural world. This issue has been identified as a part of the 'taxonomic impediment' - the lack of taxonomic information, skills, personnel and capacity

inhibiting many developing countries from implementing policies and practices of sustainable management and conservation of biodiversity. In particular, under the Convention on Biological Diversity (CBD), methods of dealing with the taxonomic impediment have been elaborated under the "Global Taxonomy Initiative". Within the Work Programme of the GTI the need to make available the contents of taxonomic literature, especially volumes like the BCA, and details of material held in collections outside the countries of origin, is highlighted several times.

Making the information held in the BCA widely available electronically, especially to Mexico and the countries of Central America, will enable a significant increase in capabilities to identify and work with the biodiversity of that region. The extraordinarily detailed and finely-engraved plates of the BCA form the basis of an unparalleled field guide to the fauna and flora. Downloaded elements, or parts captured on disk and used on non-connected machines, will enable workers in remote locations to work with the animals and plants. With the linkages between data sources planned for the project, and the addition of appropriate technology (much of it already available), the data to be captured in the project will make it possible to produce distribution maps and links to climatic and ecological data enabling predictive analysis of species ranges. The linkages put in place to extant databases, additional collection databasing, and updating the nomenclature, will provide baseline data on country biota for nations who at present would find that information virtually impossible to obtain. The range of users of the product of this project will be from the interested amateurs to taxonomists to quarantine officers.

Project Outline

- **A.** SIL will republish, in digital form, the 58 biological volumes of the 63-volume set of the ***Biologia Centrali-Americana***. Because of its scope and detail, the BCA represents an important but manageable sector of published taxonomic literature for this project. This step will provide high-resolution, bit-mapped images of the entire text and figures with a navigational device.
- **B.** SIL will rekey and code the text in a manner that will allow for integration with other major scientific datasets while still providing a scientifically useful stand-alone product.
- **C.** SIL and NMNH will engage the systematic biological community in the project to define an extensible Markup Language (XML) Document Type Description (DTD) or schema for taxonomic literature to ensure its applicability to the widest possible audiences.

Importance of the *Biologica Centrali-Americana*

According to Dr. Sandra Knapp, Research Botanist, The Natural History Museum (London), and one of the participants in the October 2001 Conference,

> "…the inaccessibility of literature relating to tropical plans and animals…leads to waste of effort on the part of taxonomists from biodiversity-rich countries, either through re-description of that which has been done before, or in just trying to get

> access to what they need. As more and more literature becomes available over the Internet, it is crucial that old literature be concomitantly more available, as the combination of old and new is essential to the good practice of taxonomy and systematics. Almost uniquely, systematics relies upon its past to define its future."[1]

The *Biologia Centrali-Americana* is a fundamental work for the study of neotropical flora and fauna and includes nearly everything known about the biological diversity of Mexico and Central America at the time of publication. The BCA was privately issued in installments between 1879 and 1915 by F. Ducane Godman and Osbert Salvin of The Natural History Museum (London). "The work consists of 63 volumes containing 1677 plates (of which more than 900 are coloured) depicting 18,587 subjects. The total number of species described is 50,263 of which 19,263 are described for the first time."[2] The contents were derived from scientific survey and explorations conducted during the latter part of the 19th and early 20th centuries. Many of the leading biologists of the time provided specimens and descriptions for the many volumes. The illustrations are, in many cases, the only images that exist of the biota of the region and as such could be compiled for use in an electronic field guide if available in a digital and portable format. The specimens described are deposited in many places including The Natural History Museum (London), Royal Botanical Gardens (Kew, UK), Missouri Botanical Garden, American Museum of Natural History, and the National Museum of Natural History. Since the BCA appeared, a few select volumes have been republished but never the entire series.

Based on research in the OCLC international bibliographic database, which contains cataloged holdings from tens of thousands of libraries, the entire 63-volume BCA is held by only eight libraries. Many other libraries hold individual volumes. Some Central American countries, whose flora and fauna are so well documented in BCA, lack a complete set; thus the BCA is not generally accessible to taxonomists working in the region.

BCA volumes in most repositories have been heavily used and are in a deteriorated state. The paper used in the production of the work was highly acidic and over time has become brittle. To protect the volumes from heavy use when possible, they are often stored in rare book vaults. Many of the volumes also contain valuable annotations and other marginalia that enhance the work with thoughts of scientists of the past 100 years. Production and retention of a high-quality digital edition will ensure survival of these characteristics and long-term preservation of the original work.

It is important to do the all the biological volumes, not just selected ones, to produce a complete picture of the biodiversity of Central America ca. 1900. In addition to a full coverage of Central American flora and fauna, the BCA set also has five volumes on the archaeology of the area. While of scholarly value, these volumes are not directly relevant to the biological thrust of this project and are not relevant for linking the BCA to

[1] Email from Dr. Sandra Knapp to Thomas Garnett, 12/3/01
[2] Prospectus, Biologia Centrali-America, p. 4.

taxonomic and specimen databases. In addition, the archaeological volumes have recently been republished by the University of Oklahoma Press and are thus easily available. To keep the project costs low, the volumes on archaeology will not be scanned and rekeyed as part of this effort.

The BCA is important to biodiversity studies and national biological surveys at such scientific centers as the Smithsonian Tropical Research Institute, INBio (Instituto Nacional de Biodiversidad) in Costa Rica, CONABIO (The National Commission for the Knowledge and Use of Biodiversity) in Mexico, and the Humboldt Institute in Colombia. However, Dr. Knapp reports:

> "….This project is not just about providing the original data for biologists in developing countries to use to empower them to adequately meet their obligations under the Convention on Biological Diversity. This is the first step in helping to remove the taxonomic impediment--and it is critical that the large museums of the world, who hold so much of the data in the form of literature and specimens for the rest of the world, be the ones to make this major attempt."[3]

Project Benefits

The full text of the BCA will be available to any researcher with an Internet connection anywhere in the world. Researchers who must now travel significant distances to use these texts will have them from their desk. Users will be able to search the BCA by species name and retrieve the relevant portions. This product, while immensely useful, will only be the beginning.

Use by researchers worldwide
The BCA is essentially an early database of biodiversity of Mexico and Central America and thus forms the basis for subsequent advances in knowledge, as well as the foundation for studying changes caused by natural or human factors (e.g., invasive species). Historians of science and other fields will find data available for their research. Libraries who lack copies of the originals will also benefit. This project will also be the keystone and model for several other major bioinformatics projects to be pursued by the leading biological repositories in the world. For example, the rekeyed text will be coded in XML so that it may be linked with records in specimen and taxonomic databases such as the Multimedia Catalogue at NMNH, W³TROPICOS at the Missouri Botanical Garden, and many others.

As a model
The BCA, while certainly one of the most comprehensive and broad-ranging, is only one of a large number of similar works that would benefit the research community through being more accessible. By providing a model of how a work like this can be successfully translated to a digital form, the stage will be set for others to engage in similar projects.

[3] Knapp, op cit.

As a basis for research and development
The digitized BCA will also engage the systematic biology community in general by providing an experimental testbed for determining the best ways of linking different, yet complementary, sets of data. A large part of the testbed will consist of defining the practices and standards needed to ensure effective crosswalks between, and links among, relevant biological data systems, including appropriate specimen images, especially of types. In particular, an XML-based standard for coding taxonomic literature will be defined as part of this project. As a recent report from the Council on Library and Information Resources stated, successful digitizing projects "will fit technologically into ("interoperate" with) broader collections so that they can be accessed through a variety of search engines and portal services far beyond a single campus."[4]

Some of this work will take place in other repositories. For example, the American Museum of Natural History has committed to fund a project to rekey and code the text of *A Gazetteer to Accompany the 'Insecta' Volumes of the 'Biologia Centrali-Americana,'* by Richard B. Selander and Patricia Vaurie.[5] This will be linked to the digitized BCA and to community-based GIS systems, such as the Alexandria Digital Library Gazetteer at the University of California, Santa Barbara.[6] In another project, the Missouri Botanical Garden plans to link the digitized BCA to the *Flora Mesoamericana* (FM), a joint effort of NHM (London), the Missouri Botanical Garden, and the Instituo de Biología of National Autonomous University of Mexico, to provide the historical context. These links are now being laboriously transcribed manually, for the FM's approximately 60,000 names and 17,000 species. The conferees at the October meeting plan to generate several additional grant proposals to other funding sources to help forward this work.

Additional follow-on projects may include production of portions of the digitized BCA on searchable cd-roms or on a chip that can be loaded into a handheld computer for use in the field.

Methodology for the Electronic BCA

Funding from the Seidell Endowment will support the following tasks:

1. SIL will correct the cataloging of the BCA in the Smithsonian Institution Research Information System (SIRIS). Because the records in SIRIS were converted from catalog cards created under differing cataloging practices and do not always reflect current locations, the records are frequently inaccurate and misleading. A cataloging contract will be employed for this step.

---

[4] ***CLIRinghouse*** (April 2002), Washington, D.C., Council on Library and Information Resources.
[5] Richard B.Selander and Patricia Vaurie. ***A Gazetteer to Accompany the "Insecta" Volumes of the "Biologia Centrali-Americana,"*** New York: American Museum of Natural History, 1962. American Museum Novitates. no. 2099.
[6] The Alexandria Digital Library is a working digital library with collections of geographically referenced materials and services for accessing those collections. The ADL Gazetteer contains 4.5 million place names. It is headquartered on the campus of the University of California at Santa Barbara and hosted by the Davidson Library (http://fat-albert.alexandria.ucsb.edu:8827/gazetteer/)

2. SIL will scan the entire set of volumes and create a digital edition of the BCA mounted on a server in NMNH. This digital edition will be a navigable, very legible organization of bit-mapped images of the pages available through the Internet to anyone with a web browser. A draft statement of work for this contract is in Appendix C.

3. The web application serving and linking the full text BCA will be implemented to complete a set of system engineering/software development tasks:

   - define and develop the web application(s) which will maintain and deliver the BCA data;
   - define and develop the metadata and database structure to support the application(s);
   - define and implement the hardware/software infrastructure to support the application;
   - define and integrate the application with existing Smithsonian applications, database management systems, and infrastructure.

   The deliverables will include: task plan, concept of operations, requirements specification, high level architecture design, rapid application development prototype of end-user interface, detailed design, coded and/or tested software and hardware components, integrated, implemented, and tested application and supporting infrastructure.

4. The text of the BCA will be rekeyed.

5. While being rekeyed, the text will be marked up in XML using a Document Type Description (DTD) developed by the American Museum of Natural History (AMNH) in a project funded by the Andrew Mellon Foundation. This will allow for full-text searching, easier on-screen reading and navigation, easy reformatting of portions of the text for multiple uses (e.g., Adobe Acrobat files) including downloading and exporting, and integrating with other systems, etc. Those portions of the text that are species descriptions will be further coded using an XML schema developed in coordination with representatives of the taxonomic community specifically for coding taxonomic literature. This will require hiring a technical consultant with extensive XML experience in the biological disciplines. This coding and the system engineering/software development in no. 3 (above) will allow linking of the BCA data to:

   a. collections information systems, such as the NMNH Multimedia Catalogue, W³TROPICOS, and other specimen databases including those held by CONABIO and the NHM (London), which contain images, especially of type specimens;

b. taxonomic databases such as the Integrated Taxonomic Information Systems (ITIS), Species 2000, the International Plant Names Index (IPNI), etc.;

c. standards-based electronic gazetteers such as the Alexandria Digital Library Gazetteer.

Additional Projects Based on the Electronic BCA

The following tasks are part of the larger project and will be contributed by or funded by other agencies:

1. The text, *A Gazetteer to Accompany the "Insecta" Volumes of the "Biologia Centrali-Americana"* will be rekeyed and coded.[7] This will be linked to community-based GIS such as the Alexandria Digital Library Gazetteer.

2. Practices and standards to ensure effective crosswalks between and links among relevant biological databases, obtain appropriate specimen images, etc., will be defined. Many other institutions will be cost-sharing these expenses with staff work and direct funding. Please see Appendix B for the October Conference proceedings and a list of participants and see Appendix A for an example of how the BCA data will be used with the *Flora Mesoamericana* project. The conference participants expect to generate several additional grant proposals to apply to other funding agencies especially in the application of the XML-based standard for coding taxonomic literature that will be defined as part of this project.

3. As the digitized BCA volumes are made available, additional funds will be sought from other sources for collaborative projects with many partners. The partners are expected to include not only the participants from the October Conference but also other organizations especially those in Mexico and Central America. In this way, the impact and importance of this project will grow and become increasingly useful in assisting those countries in environmental management and conservation, as well as being a basic tool for education and enjoyment of the natural world.

4. Long-term "archival" storage of image and coded text files will be arranged through the OCLC Digital and Preservation Co-op that SIL has joined as a charter member. OCLC is a tested and trusted third-party, which has been at the forefront in enhancing access to information resources for the last 40 years. As a charter member of the OCLC Digital and Preservation Co-op, SIL, with peer institutions, will be defining the structure, features, and policies of the OCLC Digital Archive, the first version of which is planned to open this spring.

[7] Selander and Vaurie, op cit.

Personnel

The Smithsonian Institution Libraries (SIL) will manage the BCA project, under the direction of Tom Garnett, Assistant Director for Digital Library and Information Systems. Martin Kalfatovic, Head of SIL's New Media Office, will serve as project manager and Contracting Officer's Technical Representative (COTR) for the scanning and rekeying contracts, as well as supervisor of project personnel. Sherry Kelley, Assistant Director for Technical Services and Administration will be the COTR for the cataloging clean-up project. Dawn Leaf, Director, Systems Architecture & Product Assurance, Office of the Chief Information Officer, will coordinate OCIO review of and participation in the information technology portions of this project.

SIL will hire, in a temporary staff position, an Imaging Processing Technician for the duration of the project to prepare the volumes for scanning; organize shipping and receiving workflow to maximize production and ensure a steady stream of materials for the vendor(s); inspect returned images for clarity, color, completeness, skew, etc. (inspection includes both source and derivative images); inspect CD-ROM or other long-term media storage; enter item label and image information in a management database and/or inspect vendor-produced image information created for database; and compile statistics for projects and individual work.

SIL will form a scientific advisory committee to assist in guiding the project and to ensure it will be of maximum utility to the scientific community. The committee will meet at least quarterly during the project to review progress.

In addition, the Smithsonian plans to establish working groups composed of the participants at the October conference to advance related projects and reach consensus for common standards-based approaches to these issues.