



## THE BIODIVERSITY HERITAGE LIBRARY: SHARING BIODIVERSITY LITERATURE WITH THE WORLD

**Dr. Nancy E. Gwinn**

Director, Smithsonian Institution Libraries

**Constance Rinaldo**

Librarian, Ernst Mayr Library and Archives,  
Museum of Comparative Zoology  
Harvard University

**Meeting:**

**Simultaneous**

**Interpretation:**

**109. Science and Technology Libraries**

English, Arabic, Chinese, French, German, Russian and Spanish

---

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL

10-14 August 2008, Québec, Canada

<http://www.ifla.org/IV/ifla74/index.htm>

---

### ABSTRACT

*Ten major natural history museum libraries, botanical libraries, and research institutions in the United Kingdom and the United States joined in 2005 to develop a strategy and operational plan to digitize the published literature of biodiversity held in their respective collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons.” Headquartered at the Smithsonian Institution Libraries, the Biodiversity Heritage Library (BHL) has been embraced as one of the four cornerstones of the Encyclopedia of Life, an unprecedented global effort to document all 1.8 million named species of animals, plants, and other forms of life on earth. Major funding for the project has come from the John D. and Catherine T. MacArthur Foundation, the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation and the Richard Lounsbury Foundation. This paper will provide an overview of the BHL and its potential impact on biodiversity research, describe the BHL portal and its innovative search services, and provide a case study of the process from one of the members: the Museum of Comparative Zoology at Harvard University.*

## INTRODUCTION

According to a U.S government report, biodiversity, a term coined by internationally renowned scientist E.O. Wilson as a shortened form of “biological diversity,” refers to the variety and variability among living organisms and the ecological complexes in which they occur.<sup>1</sup> The basic sciences to identify biological diversity are known as taxonomy, the theory and practice of describing, naming, and classifying all plants, animals, and microorganisms of the world, and systematics, the classification of living things into groups based on their evolutionary origins. Research in these disciplines is most often conducted in natural history museums and botanical gardens in the developed world, which contain rich and extensive collections of flora and fauna, as well as large libraries of the literature that document them. Taxonomists have barely scratched the surface of known species (Godfray 2007; Wheeler 2008). It is estimated that there are between 5 and 30 million species on the planet, yet only 1.8 million have been identified and described.

In 1992, 150 government leaders signed the Convention on Biological Diversity, which recognized the crucial role of taxonomy in promoting sustainable development. For several years, taxonomists have recognized the need to speed up their work, before expanding populations, environmental calamities, and economic development reduce the wealth of existing species. Their work has value well beyond the act of identification; it has wide use and economic impact for a broad range of applications in agriculture, biodiversity conservation, protected area management, control of invasive species, forestry, plant breeding, disease control, and trade in natural products, including pharmaceuticals (Wheeler 2008). Yet there are severe obstacles to progress, what the field has termed the “taxonomic impediment.” Taxonomists largely carry out

---

<sup>1</sup> U.S. Congress. Office of Technology Assessment. 1987. *Technologies To Maintain Biological Diversity*, OTA-F-330 (Washington, DC: U.S. Government Printing Office): 313.

their work among the specimen and literature collections in the industrialized nations. There are few taxonomists to collect and work on the millions of undescribed species in biodiversity-rich but economically poorer countries.<sup>2</sup> And for all taxonomists, access to the relevant literature can be a costly, time-consuming process (Godfray 2007; Minelli 2003).

More than any other science, the domain of systematic biology is utterly dependent on the historical literature of published descriptions of species; publication in print still determines the legitimacy of naming and credit for new discoveries. Alessandro Minelli writes:

According to current practice, any serious monographic work about a given taxon should include a careful consideration of all previous literature dealing with at least one species belonging to that taxon, regardless of whether it was published last year or in the late 18<sup>th</sup> century, whether it is written in English or in Spanish, in Russian, or in Latin. If these works include the description of (the then) new taxa, or have an impact on its nomenclature, they cannot be ignored.<sup>3</sup>

Minelli describes taxonomic papers as “legal” documents as well as scientific ones, because they describe new species according to rules of distinct international codes.<sup>4</sup> The codes exist to ensure that all taxonomists adhere to principles of priority, which resolves problems caused by the use of homonyms or synonyms in naming. Consequently, taxonomists must consult all relevant literature from Linnaeus onwards to ensure a sound basis for their work.

Technology and the Internet finally provided a way to dissolving the taxonomic impediment, at least in part, through scanning of both the literature and specimen collections so

---

<sup>2</sup> “Global Taxonomy Initiative, Convention on Biological Diversity, <http://www.cbd.int/gti/problem.shtml>

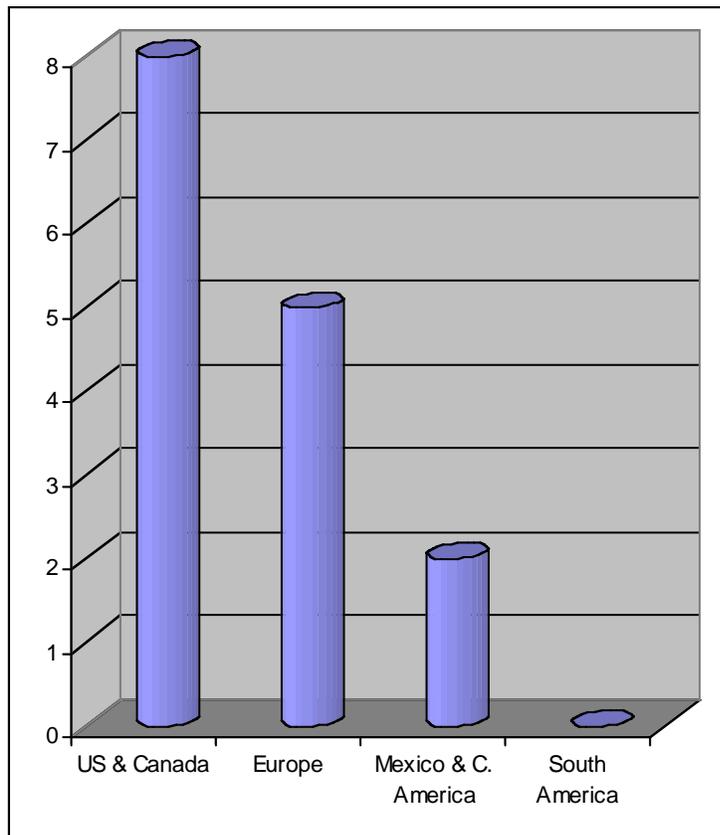
<sup>3</sup> Minelli, A. 2003. “The Status of Taxonomic Literature,” *Trends in Ecology and Evolution* 18(2): 75.

<sup>4</sup> These are the *International Code of Zoological Nomenclature*, the *International Code of Botanical Nomenclature* and the *International Code of Nomenclature of Bacteria*.

they can be shared with the global scientific community (Godfray 2007). There are over 5.4 million volumes on biodiversity dating back to 1469: 800,000 monographs and 40,000 journal titles. Fifty percent were published before 1923 and are in the public domain in the United States. It is a big job.

## PLANNING BEGINS

In 2003, the Smithsonian Institution hosted a conference of biologists to discuss what was needed to improve the efficiency of biological research. The greatest obstacle, experts determined, was access to the historical literature. Those of us in natural history museum and botanical garden libraries know that for years, researchers have traveled to use our collections, often spending most of their time standing in front of photocopy machines to collect as much as



*Figure 1: Distribution of copies of the Biologia Centrali-Americana; the copies in Central America are located in the Earl Tupper Library, Smithsonian Tropical Research Institute, Republic of Panama, one of the Smithsonian Libraries' twenty branch libraries. Courtesy, Martin Kalfatovic.*

possible before they returned home. Digital technology and the Internet offered a solution; if the literature was scanned and made searchable on the web, researchers could gain access from wherever they were in the world. As a result of the conference, the Smithsonian provided funding for the Smithsonian Libraries to digitize the *Biologia Centrali-Americana*, a 63-volume work published from 1879-1915, which remains the seminal work on the flora and fauna of

Central America.<sup>5</sup> The funding also supported initial work toward developing a system of automatic coding of the scientific names of species found in the work to improve accessibility for taxonomists.

In 2005 at the Natural History Museum in London, the Alfred P. Sloan Foundation funded an international symposium called *Library and Laboratory: the Marriage of Research, Data, and Taxonomic Literature*. The eighty biologists, librarians, and computer scientists who attended again identified the lack of access to the published literature of biodiversity as one of the principal obstacles to efficient and productive research (Moritz 2005).

In May 2005, representatives of several major natural history and botanical libraries met at the Smithsonian National Museum of Natural History in Washington, D.C. With funding provided by the Smithsonian Institution, the goal of the meeting was to develop a strategy and operational plan to digitize the published literature of biodiversity held in their respective collections and to make that literature available for open access and responsible use as a part of a global “biodiversity commons” (Moritz 2002). Two years later, the directors of the libraries of the American Museum of Natural History, Harvard University Botany Libraries, Harvard University Ernst Mayr Library of the Museum of Comparative Zoology, Missouri Botanical Garden, Natural History Museum in London, New York Botanical Garden, the Royal Botanic Gardens, Kew, Field Museum of Natural History in Chicago, Marine Biological Laboratory/Woods Hole Oceanographic Institution Library and the Smithsonian Institution Libraries agreed to a Memorandum of Agreement that established the Biodiversity Heritage

---

<sup>5</sup> The *Biologia Centrali-Americana* (BCA) is a fundamental work for the study of neotropical flora and fauna. It includes nearly everything known about the biological diversity of Mexico and Central America at the time of publication (1879-1915). The original work consists of 58 biological volumes containing 1284 plates illustrating 18,587 subjects. A total of 49,392 species are described, 19,263 for the first time. Many of the illustrations and descriptions are the only ones that exist of the biota of the region.

Library.<sup>6</sup> The Smithsonian Libraries agreed to host the BHL Secretariat and provided from its senior management staff the first Program Director, Thomas Garnett.

Another development was brewing at the same time, the Encyclopedia of Life.<sup>7</sup> This is an ambitious, even audacious, collaborative global project to document authoritatively the 1.8 million known species of animals, plants, and other forms of life and to create web pages on the Internet for each one. Each page is just an entry point, suitable for the general public but with linked pages pointing to more specialized data for researchers. The Encyclopedia embraced the Biodiversity Heritage Library as one of its four cornerstones, to bring much of the historical literature about a species to the relevant web page. This was critical for the project, because through the Encyclopedia of Life, the BHL received a first increment of \$3 million from the John D. and Catherine T. MacArthur Foundation. Additional funds have come from the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, the Richard Lounsbury Foundation and individual BHL member institutions.

## **WHY DO THIS NOW?**

The ten BHL member libraries have over two million volumes of biodiversity literature collected for over 200 years to support scientists and students throughout the world. Clearly these libraries together hold a substantial part of the world's published knowledge on biological diversity. While there are several mass digitization projects at major research libraries here and abroad, none have the discipline-specific focus of the BHL partner institutions and may fail to capture significant elements of this biodiversity legacy. Much of the biodiversity literature is highly specialized and often not duplicated even in broad university research collections. Nor

---

<sup>6</sup> "About the Biodiversity Heritage Library," <http://www.biodiversitylibrary.org/About.aspx>

<sup>7</sup> Encyclopedia of Life <http://www.eol.org>

are all of them allowing open access to the digitized publications. With its innovative search strategies, the BHL is intended to be “one-stop shopping” for those needing to consult biodiversity literature.

Costs of scanning have fallen considerable and in a high-production mode, the Internet Archive is projecting a low basic cost of 10 cents a page. The biodiversity literature is a tractable, well-defined scientific domain and has extreme longevity—current taxonomic literature often relies on texts and specimens that are more than 100 years old. (Godfray 2007; Minelli 2003) In addition, the BHL supports the Global Biodiversity Information Facility (GBIF) and other international biodiversity initiatives<sup>8</sup> (Speers and Edwards 2008). The benefits are clear: taxonomists and other scientists will have access to the biodiversity literature – globally, scientists and citizens in the developing world will finally have easy access to the historical literature. This clearly advances the objectives of the international Convention on Biological Diversity.<sup>9</sup>

## **GETTING STARTED**

The BHL members selected the Internet Archive (IA) to provide the scanning services and also to archive the resulting digital files. The Internet Archive both shares the BHL mission to provide open access to the literature and has the capacity to do mass scanning in a high production mode at reasonable cost. The Internet Archive began digitizing for BHL partners in early 2007. BHL members are working with three of the six ten-station Internet Archive digitization centers (Boston, Washington, D.C., and New York), as well as with smaller centers in the Smithsonian Institution, the University of Illinois, Urbana-Champaign, and the Natural

---

<sup>8</sup> Global Biodiversity Information Facility, <http://www.gbif.org>.

<sup>9</sup> Convention on Biological Diversity, <http://www.cbd.int/>

History Museum, London. However, the partners are doing much more than simply scanning volumes.

Missouri Botanical Garden staff established the Biodiversity Heritage Library portal as an innovative research environment with the ambitious goal of transforming the nature of scientific inquiry, as well as vastly accelerating research in life sciences and conservation.<sup>10</sup> The portal provides access to both digitized images (JPEG 2000, PDF, and JPEG) and Optical Character Recognition (OCR) text of the literature, but that is only the start. It also employs an array of taxonomically intelligent services designed to overcome the problem of common name versus scientific name and changes of names over time. This scientific reference system for investigating scientific literature offers a model that reflects, and also serves to amplify, scientists' approaches to, and use of, the body of natural history literature.

## **WHAT IS TAXONOMIC INTELLIGENCE?**

Scientists use scientific names to find information about organisms. One organism can have many scientific names over time or multiple common names depending on language or region. Additionally, one name might refer to multiple organisms. Thus it can be difficult to retrieve information about an organism even if the current scientific name is known. This problem was addressed by an international project called uBio for Universal Biological Indexer and Organizer. uBio is composed of the Taxonomic Name Server (TNS), which acts as a name thesaurus; NameBank, a repository of over 10.7 million recorded biological names and identifiers that link those names together; and ClassificationBank, which stores multiple

---

<sup>10</sup> Biodiversity Heritage Library, <http://www.biodiversitylibrary.org>

classifications and taxonomic concepts.<sup>11</sup> BHL uses [TaxonFinder](#), a taxonomic intelligence algorithm developed by the collaborators at uBio, to compare the OCR texts with NameBank and identify likely scientific names. Once fully integrated, a researcher will then be able to search the BHL collection using any form of an organism’s name (Leary, et al. 2008). This will also allow users searching the Encyclopedia of Life web pages to draw in the literature related to the species.



Figure 2: A page in the Biodiversity Heritage Library showing the list of scientific names (bottom left lower corner) extracted from the OCR text.

After the Internet Archive scans the volumes, the BHL portal ingests MARXML metadata and low-resolution JPEG files. High resolution JPEG2000 files are retrieved on the fly from the Internet Archive when requested by a user and decoded at the portal for viewing via a web-browser. OCR text is also sent on the fly to uBio for name extraction via taxonomic

<sup>11</sup> “About the project,” <http://www.ubio.org>

intelligence and displayed in real time with the page image. The BHL portal serves the image and text files; displays volume, part and piece metadata; and applies Globally Unique Identifiers (GUIDs) for linking to other taxonomic services. Persistent URLs allow linking at bibliographic record, volume, and page levels in BHL. The BHL technical staff at the Missouri Botanical Garden staff are constantly improving the portal's presentation of results and adding features such as geocoding the Library of Congress Subject Headings using the Google Maps Application Programming Interface and "discovered" bibliographies of previously un-indexed species references.(Freeland, et al. 2008).

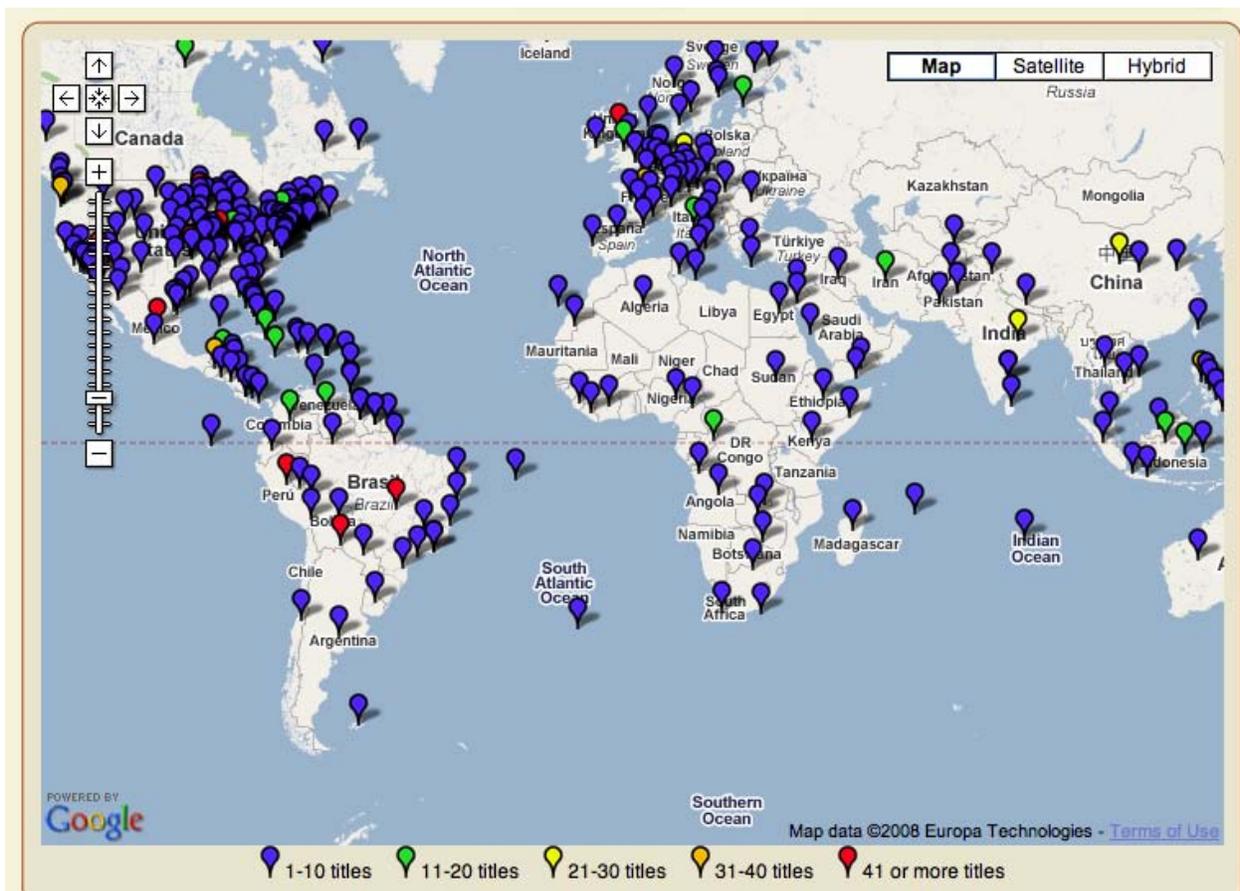


Figure 3: The Library of Congress Subject Headings of the volumes in the BHL are geocoded and then mapped using the Google Maps API

## **BUT WHAT ABOUT COPYRIGHT?**

In general, the BHL project attempts to keep copyright infringement risk low by tackling the public domain literature first, seeking permissions for digitization, negotiating alternative agreements and moving on when none of these tactics applies. BHL has an opt-in copyright model. The BHL Program Director has opened negotiations with a variety of publishers from small, learned societies to large commercial organizations. As of April, 2008, the BHL has obtained permissions to digitize forty-nine titles from museum and small society publishers. The BHL will digitize the entire run of the publications to the most recent issues, as per the negotiated permissions, and mount them on the BHL portal at no cost to the societies. The files can be reused by the society for its own purposes.<sup>12</sup> The BHL will take responsibility for long-term sustainability of the scanned material. Some aggregators and commercial publishers have expressed interest in alternative agreements, such as providing metadata and OCR files for indexing using taxonomic intelligence tools.<sup>13</sup> Discussions are underway with the *Zoological Record* and a collaboration with BHL is possible.

## **BHL GOVERNANCE**

The Biodiversity Heritage Library is not incorporated and thus is not a legal entity. Through the signed Memoranda of Agreements, each member institution has committed to the collaborative effort to build the Biodiversity Heritage Library and to search for additional funding to support it. The directors of the member libraries form the Institutional Council, which meets at least annually to review progress, discuss current issues and elect officers. An Executive Committee meets weekly by conference call and consists of the BHL Program Director, the Institutional Council Chair, Vice-Chair, Secretary and the Technical Director. As

---

<sup>12</sup> "Can I Use Your Images," <http://www.biodiversitylibrary.org/Copyright.aspx>.

<sup>13</sup> "Tools," <http://biodiversitylibrary.org/Tools.aspx>

members have implemented the scanning process, practical issues and obstacles have appeared. The case study of the Ernst Mayr Library at Harvard's Museum of Comparative Zoology illustrates some of these.

## **ERNST MAYR LIBRARY CASE STUDY**

The costs of scanning, while low on a per-page basis, are high over the course of the entire project. Thus it was essential to identify ways to minimize duplication of scanning. The first step taken was to purchase the OCLC Collection Analysis tool and add all member library records.<sup>14</sup> The Smithsonian Institution paid for the first year's subscription. It took more than a year to get this tool up and running, partly because not all libraries were members initially of OCLC, and, in the case of Harvard University, the records of the botany and zoology libraries could not be separated from those of the other Harvard libraries. The OCLC Collection Analysis tool enabled a broad look at institutional collection strengths and allowed the group to estimate the numbers of public domain pages that could be scanned more or less immediately and to identify subjects with unique titles in the member libraries. Non-BHL mass scanning projects have some overlap with the BHL. To minimize duplication, library staff review other projects; if the material is open access and easily ingested, in some cases, BHL members may alter scanning plans.

With input from the EOL Steering Committee, BHL members chose serial publications as the first priority, because the serial literature is critical to the scientists who make up the primary audience for the first release of the BHL. The systems office of the Natural History Museum of London developed a tool to avoid scanning duplication among the member institutions. A mashup database was generated that included online catalog records of all the

---

<sup>14</sup> "WorldCat Collection Analysis," [www.oclc.org/collectionanalysis/default.htm](http://www.oclc.org/collectionanalysis/default.htm)

serials in all the member institutions with a checkbox for selection; thus a serials bid list was born. As titles are identified for scanning, an institution will access this database to make sure no one else has bid on the item. If there are no bids, or bids cover only part of a serials run, the institution is free to claim the title, or at least the parts of the title that have not been bid, for scanning. The MBL/WHOI Library has built a monograph de-duping tool that shows what monographs have been scanned; there is no bidding process for monographs as yet.

BHL members established a wiki early in the project to facilitate communication among the widely dispersed libraries. The wiki has been used to maintain minutes of conference calls and meetings, post presentations and procedures, and coordinate technical questions, quality discussions and collections work. It became clear in March 2008 that member staff managing the day-to-day planning and processing workflows required better communication. Staff needed to compare their experiences more directly to ensure efficient workflows and to avoid “reinventing the wheel.” Led by Suzanne Pilsk from the Smithsonian Libraries, staff now participate in two or more monthly conference calls. These calls have become critical to developing and managing efficient and coordinated workflows for the BHL project. Staff in the member libraries discovered that everyone was experiencing frustration with various workflow problems and were surprised at the overwhelming amount of work to be done.

James Hanken, the Director of the Museum of Comparative Zoology set aside funding for the next five years to digitize the Ernst Mayr Library collections for the Biodiversity Heritage Library. The funds are for digitization, processing and moving costs. With this funding, the library hired two Simmons College Library and Information Science students as interns to prepare the collection for scanning. The primary role of the interns is to barcode, pull and cross-check library materials for suitability for scanning and to flag foldouts.

In the Ernst Mayr Library, the Head of Technical Services uses reports generated from the integrated library system (ILS) to create “picklists” of items to pull for scanning. Picklists are lists of items available to scan, based on date criteria, in shelf-list order. The picklists make it easier for the interns and other staff to go to the shelves and pull exactly the right items. Picklists are edited upon checking the holdings of BHL partner institutions using tools described above to minimize duplication of scanning effort. The Head of Technical Services then checks items pulled based on the picklists, generates a final packing list for each shipment to the scanning center, checks books leaving and returning against the packing list, and oversees the work of the Simmons interns. A database provides an inventory of all items scanned for the Ernst Mayr Library. Items rejected, either by the Library or the scanning center, are tracked via the integrated library system. The Special Collections Librarian provides subject specific (e.g. Herpetology, Ichthyology) lists of monographs in the Special Collections and pulls appropriate items for barcoding and shipment to the scanning center. The Circulation Supervisor also pulls items and then reshelves them when they return.



*Figure 4: One of the Internet Archive scanning stations at the Boston Public Library.*

Currently at least 2.5 FTE (including Constance Rinaldo's time as BHL Secretary) are devoted to the project. Additionally, the Harvard University Systems Office worked with the Internet Archive to ensure smooth Z39.50 connections to help with project planning and glitches and to add URLs to HOLLIS, the Harvard University's online catalog on return of the digitized material. The Z39.50 connection enables IA to pull metadata from the Harvard catalog to populate the IA records. Permission by non-Harvard entities to use Z39.50 access is granted only on a case-by-case basis; gaining this permission took more than 6 months. Ultimately, it is important to Harvard that Harvard users have immediate access to the Harvard collections that are available electronically. Thus the final step is to provide access to digitized material through Harvard's portal. This means that the persistent URL for Harvard-generated digital objects must

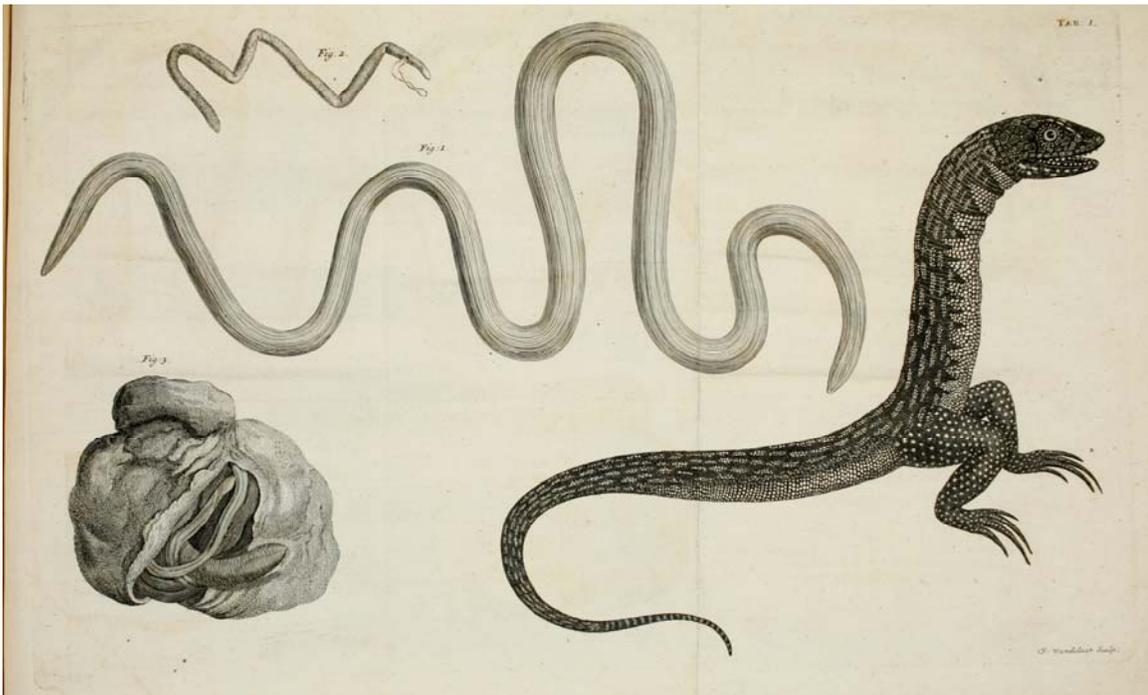
be made available in HOLLIS.. At this time, persistent URLs generated by the BHL portal must be entered by hand, but the Harvard University Office of Information Systems is working with the Ernst Mayr Library to develop an automated solution.

For the Ernst Mayr Library, the Office of General Council at Harvard University has determined that classes of texts that can be digitized at low risk include Harvard University, Museum of Comparative Zoology (MCZ) publications, pre-1923 American copyrighted materials and pre-1908 non-American copyrighted materials. Additionally, where permission has been granted, other museum and society journals can be scanned.

Developing a contract has been a lengthy (two-year) process with back and forth between Harvard and Internet Archive lawyers. Although scanning has been under way since December, 2007, the contract has just been accepted. The Ernst Mayr Library hired a library moving company to shuttle books between the Ernst Mayr Library in Cambridge, MA and the Boston Library Consortium Scanning Center in the Boston Public Library. Access to the Ernst Mayr Library is challenging for the movers because of circuitous routes due to new construction and the difficulties in navigating around the old building.

Many of the early publications, particularly those from the MCZ have fold-outs or are larger than average size. Initially, the Internet Archive did not have a fold-out solution and staff had to pull and mark any volume with fold-outs, and these were not scanned. Other issues that cause rejection or questioning of items for scanning include tight margins, size, brittleness and poor binding. IA personnel in the Boston scanning center have been very flexible and are willing to scan unusual items and provide test scan samples so that the best decision for the Ernst Mayr Library materials can be made. A fold-out solution is being tested in Boston and although there is still a size limit (18" by 24"), it is an exciting improvement. The foldouts scanned by IA

vary from being superior to the original to as good as the original to being inferior to the original but adequate for display. To see the best quality foldouts, the JPEG2000 image must be viewed. The pdf and other views of the foldouts are not adequate in most cases. Fewer books are rejected now that there is a budding foldout solution but scanning the foldouts increases the time needed to scan an item and thus the cost. The Ernst Mayr Library is currently sending approximately 200 volumes every other week to the IA scanning facility at the Boston Public Library.



*Figure 5: A sample foldout from Klein, Jacob Theodor. 1755. Tentamen herpetologiae. Leidae & Gottingae: Apud Eliam Luzac, Jun. The original foldout was 10" by 15".*

## **RESULTS TO DATE**

As of April 20, 2008, there are more than 3.5 million pages, contained in nearly 4,000 titles (more than 9,000 volumes) accessible through the BHL portal. The project has demonstrated that:

- the concept of mass scanning of general collections is possible,
- there are high levels of OCR accuracy in late 19<sup>th</sup> and 20<sup>th</sup> century printing,

- the taxonomic intelligence (species name finding) across millions of pages against nearly 11 million names in Name Bank is highly effective, and
- administratively separate and geographically disparate institutions can collaborate on a complex, multi-level project and achieve concrete results in a specific knowledge domain.

Most of the literature is in the developed world, in the Northern Hemisphere, but as mentioned earlier, most of the biodiversity is in the developing world and the Southern Hemisphere. One of the most desirable outcomes of the project is to repatriate literature to the originating countries of publication by making it accessible to anyone with an Internet connection. That this is already working is evident from the following message that the Smithsonian Libraries received from a researcher in a natural history museum in Peru:

My deepest gratitude for allowing me access to the digital version of the very rare "Bulletin des Séances de la Société Entomologique de France". It has been very important for my work on the database of the names of the butterflies of the world to be able to consult at leisure this series, which is held by extremely few libraries in the world. I cannot stress enough the importance of having access to electronic versions of the literature, especially to us researchers who cannot benefit from well-endowed institutional libraries. . . . I only wish that there were many more such electronic resources. Please keep up the excellent work!

and another from Hawai'i:

Aloha. I live on The Big Island of Hawai'i, a \$300.00 plane ride away from Honolulu and the Bishop Museum. Even when I can make it to the Museum (where I study the Hawaiian Bird Skins), they do not have every single bird (moho apicalis, the Oahu moho is missing)...I have been looking for this text for over TWENTY YEARS. Mahalo nui loa for all your hard work. Reading these pages mean so much to me and many others. . . . I cannot thank you enough, nor stress the importance of your website enough. Thank you for putting these items on the web, and in such a findable manner.

The Encyclopedia of Life and other informatics projects are data mining the BHL for information located in the legacy taxonomic literature. A number of learned societies have been attracted by the opt-in copyright model and have given permissions for digitization of current

materials. Additionally the project has generated excitement in the international community and many opportunities to develop new partnerships. There has also been support from traditional bibliophile and scientific audiences

## **WHERE DO WE GO FROM HERE?**

As the case study has shown, there are problems to be solved related to scanning and the types of volumes that form the heritage literature of biodiversity. Ultimately, the project must provide “article-level” analysis of serials, which may require some adaptation of existing bibliographic indices of biodiversity literature, as well as the development of automated tools for structural analysis. This development may be combined with the creation of the “union catalog” to provide “one-stop” access to the literature.

The Biodiversity Heritage Library partners are currently Anglo-American centered, though many of the collections contain extensive non-English language material. A key goal is to develop global partnerships and include a global community. The consortium plans eventually to expand and bring in more partners, especially in Europe, Asia, and the developing world. The best way to incorporate more languages is to partner with other countries and have the BHL served from multiple nodes in multiple languages. The BHL is working to engage European partners through projects such as the European Distributed Institute of Technology (EDIT)<sup>15</sup> and Synthesis of Systematic Resources (SYNTHEsys).<sup>16</sup> EDIT consists of 27 European, North

---

<sup>15</sup> “The European Distributed Institute of Taxonomy, EDIT, is the collective answer of 27 leading European, North American and Russian institutions to a call of the European Commission, issued in 2004, for a network in ‘Taxonomy for biodiversity and Ecosystem Research.’ This project started on 1 March 2006 and will last 5 years.” <http://www.e-taxonomy.eu>

<sup>16</sup> SYNTHEsys is “the European Union-funded Integrated Infrastructure Initiative grant. This five year project, which began in February 2004, comprises 20 European natural history museums and botanic gardens [and] aims to create an integrated European infrastructure for researchers in the natural sciences.” <http://www.synthesys.info/index.htm>

American and Russian institutions whose goal is to overcome the taxonomic impediment by building taxonomic expertise through training and information provision. The goal of the SYNTHESYS project is to develop a coordinated European infrastructure for researchers in the natural sciences. The BHL hopes to leverage these projects into partnerships with other countries and thus develop an infrastructure that has a non-English component. Graham Higley, the Chair of the Institutional Council of BHL and leader of the SYNTHESYS project and the BHL Program Director, Tom Garnett, have worked tirelessly to ensure that European, African, South American and Asian countries are included in the literature discussions. Informative discussions have taken place already with the Chinese Academy of Sciences, Brazil, European Digital Library and the Museum fur Naturkunde der Humboldt-Universitat zu Berlin. The Australian Government is likely to fund scanning as part of the Atlas of Australian Life. The BHL will also continue to work with commercial publishers for fair and equitable use of their publications.

Linkages to other data types (e.g., molecular, morphological, phenotype) are key to the overall plan. It will also be necessary to get equal cost efficiencies and speed for special collections and items with fold-outs or that are oversized. Optical Character Recognition (OCR) must improve for older publications and those in non-Roman scripts. It is likely that audiences will be expanded through social networking tools and repurposing content for new audiences. The BHL was developed for scientists, but the audience will broaden as more tools become available.

The underlying architecture of BHL needs strengthening. The BHL is moving from the .Net application environment to the Fedora Commons architecture.<sup>17</sup> Fedora Commons provides an open source architecture that can manage many forms of digital content. It will allow for an

---

<sup>17</sup> “About Fedora Commons,” <http://www.fedora-commons.org/about>

open framework for the description, storage, and retrieval of digitized materials in the different ways needed by scientists and librarians. For instance, *Biologia Centralia Americana* has been described as a single title with 63 volumes or as 21 titles, each with multiple volumes. Fedora Commons can accommodate such variations and it has the potential to be a persistent, sustainable architecture for the Biodiversity Heritage Library.

The BHL partnership is working with the global taxonomic community, publishers, the Internet Archive and other organizations, to ensure that the biodiversity heritage literature is available to all, from the student seeking information on insects in the garden to scientists investigating endangered species in Brazil. It is a grand vision. But if the BHL is able to support biodiversity research by helping to eliminate one of the largest impediments to the identification of the world's living organisms, it will be counted a great success.

## REFERENCES & LINKS

- Biodiversity Heritage Library: <http://www.biodiversitylibrary.org> (accessed April 23, 2008)
- Biodiversity Heritage Library Prospectus: <http://www.sil.si.edu/bhl/supportdocuments/BHLP-prospectus10-05.pdf> (accessed April 23, 2008)
- Convention on Biological Diversity <http://www.cbd.int/>
- Encyclopedia of Life: <http://www.eol.org> (accessed April 23, 2008)
- European Distributed Institute of Technology (EDIT) <http://www.e-taxonomy.eu/>
- Freeland, C., M. Kalfatovic, J. Paige, and M. Crozier. 2008. "Geocoding LCSH in the Biodiversity Heritage Library," *The Code4Lib Journal*, Issue 2. 11 pp.
- Global Biodiversity Information Facility (GBIF) <http://www.gbif.org/>.
- Godfray, H.C.J., B.R. Clark, I.J. Kitching, S.J. Mayo, M.J. Scoble. 2007. "The Web and the Structure of Taxonomy," *Systematic Biology*, 56 (6): 943-955.
- Leary, P. R, D. P. Remsen, C. N. Norton, D.J. Patterson and I.N. Sarkar. 2008. "uBioRSS: Tracking Taxonomic Literature Using RSS," *Bioinformatics* 23(11): 1434-1436.
- Minelli, A. 2003. "The Status of Taxonomic Literature," *Trends in Ecology and Evolution* 18(2): 75-78.
- Moritz, Thomas. 2002. "Building the Biodiversity Commons" *D-Lib Magazine* 8 (6):
- Moritz, T. 2005. "Macro-economic Case for Open Access." In Proceedings of a Conference in Association with an International Conference on the Barcode of Life, organised by Graham Higley and Chris Lyal (Natural History Museum, London), Anna Weitzman and Tom Garnett (Smithsonian Institution, Washington), and Tom Moritz (American Museum of Natural History, New York).
- OCLC Collection Analysis Tool <http://www.oclc.org/collectionanalysis/default.htm>
- Speers, L, and J. Edwards. 2008. "International Infrastructure of Enabling the New Taxonomy: the Role of the Global Biodiversity Information Facility (GBIF). In *The New Taxonomy*, The Systematics Association Special Volume Series 76 edited by Quentin D. Wheeler, Boca Raton, FL.: CRC Press, Taylor & Francis Group, pp 87-94.
- Synthesis of Systematic Resources (SYNTHEsys) <http://www.synthesys.info/index.htm>
- uBio <http://www.ubio.org/>
- U.S. Congress. Office of Technology Assessment. 1987. *Technologies To Maintain Biological Diversity*, OTA-F-330 (Washington, DC: U.S. Government Printing Office)
- Wheeler, Q. 2008. "Toward a New Taxonomy." In *The New Taxonomy*, The Systematics Association Special Volume Series 76 edited by Quentin D. Wheeler, Boca Raton, FL.: CRC Press, Taylor & Francis Group, pp. 1-18
- Wheeler, Q. "Taxonomic Shock and Awe." In *The New Taxonomy*, The Systematics Association Special Volume Series 76 edited by Quentin D. Wheeler, Boca Raton, FL.: CRC Press, Taylor & Francis Group