THE BIODIVERSITY HERITAGE LIBRARY: AN EXPANDING INTERNATIONAL COLLABORATION

Constance Rinaldo
Librarian of the Ernst Mayr Library
Museum of Comparative Zoology
26 Oxford Street
Harvard University
Cambridge, MA 02138 USA

Catherine Norton
MBLWHOI Library
Marine Biological Laboratory
Woods Hole Oceanographic Institution
Woods Hole, MA USA

Abstract: The Biodiversity Heritage Library (BHL; http://www.biodiversitylibrary.org/), one of the cornerstones of the Encyclopedia of Life (eol.org), now contains more than 26 million digitized pages for almost 37,000 titles of the published literature of biodiversity held in the collections of major natural history libraries. This paper describes the development of international partnerships and expanded collaborations with scientists as well as the taxonomic tools that are integral to the BHL. BHL-Europe has now formed, China has signed an MOU with EOL and is poised to sign and MOU with BHL. Other countries and projects are ready to augment the available literature and provide redundant repositories and mirror sites. New tools such as a PDF-generator, article repository, updated search interface and social networking tools are available.

Keywords: Biodiversity Heritage Library, biodiversity, Encyclopedia of Life, E.O. Wilson, taxonomic intelligence, digitized biodiversity literature, open access

The Biodiversity Heritage Library (BHL; Figure 1) began as a consortium of natural history museum, botanical garden libraries & research institutions in the United States and London. As the idea for BHL was brewing, E.O. Wilson, the Harvard Pulitzer Prize biodiversity scientist, was encouraging the development of the Encyclopedia of Life (EOL; http://www.eol.org/ Wilson, 2007), the project to produce a web site for every species. BHL became one of the key components of the EOL.

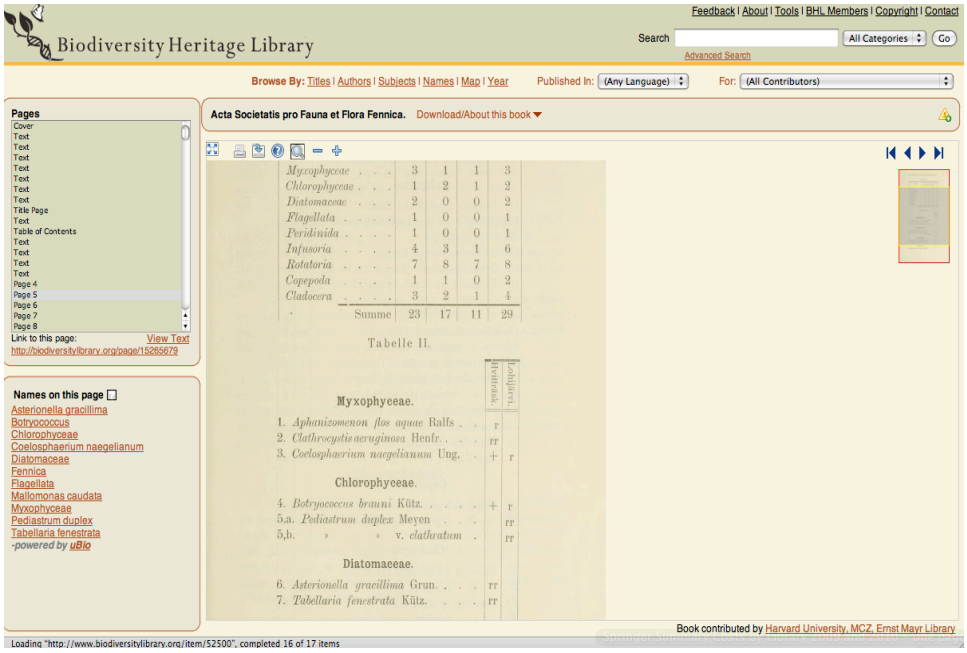

Figure 1: Old interface--not very flexible; search available for title, MARC subject, species name and author. Note the list of taxonomic names derived from the OCR (taxonomic intelligence).

The mission of the BHL is to provide open access to the literature of biodiversity. The domain of the biodiversity literature can be defined: approximately 5.4 million volumes (about 800,000 monographs and 40,000 journal titles) dating back to 1469. At least 50% of the literature was published before 1923 and thus potentially in the public domain in the United States (numbers courtesy of the Smithsonian Institution). These numbers are important because they show that there is a large body of literature (estimates suggest as much as 120 million pages) that can be digitized for open access and made freely available to all. This wealth of knowledge has historically been available only to those

few who can gain direct access to the great physical collections in these museums mainly in North American and Europe (Gwinn and Rinaldo 2009, Rinaldo 2009). Much of the printed biodiversity literature is rare, precluding easy access. No single natural history museum or botanical garden library holds the complete corpus of biodiversity literature, thus the implementation of the BHL demands wide collaboration among different types of institutions.

When the BHL was first conceived, scientists who study taxonomy were the primary audience. Thus BHL had to be more than a collection of digitized books. The information architecture includes algorithms (http://www.ubio.org/ Norton 2008) to use the OCR text to detect taxonomic names from within the scanned literature. Known as "taxonomic intelligence" this enables researchers to access the individual pages of the collection not only by title, author, and subject but also by scientific names (Figure 1). Taxonomists require access to all of the historical literature of published species descriptions in their specialty, thus the early focus on public domain literature fit the needs of this group.

In just 4 years, the BHL has grown to include technology partners, service providers, museums and research institutes all over the world. The partner libraries collectively hold a substantial part of the world's published knowledge on biological diversity. The newest members of the BHL include BHL-Europe through eContent plus funding for more than 20 museums, libraries and research institutes across Europe, China through an MOU with the Chinese Academy of Sciences and Australia through a pending MOU with the Atlas of Living Australia. More partnerships are on the horizon with active participation for a Latin American BHL through BIREME in Brazil and an Arab Language BHL through an MOU with the Library of Alexandria to be reviewed at the Pan-Arab Biodiversity Conference in Cairo, Egypt in December 2009. These projects will work together with the original BHL members to share content, protocols, services, and digital preservation practices. By collaborating, each partner institution gains the digital collection of the others, enabling the libraries to give users materials that they otherwise would not have been able to provide.

The Internet Archive (http://www.archive.org/) is the scanning partner for the BHL. The Internet Archive digitizes the materials inexpensively and provides free storage, optical character recognition (OCR) and derivative products such as pdfs and jpeg 2000. Natural history content from other Internet Archive library contributors including the California Digital Library has recently been added to the BHL, nearly doubling the content. Partnerships with professional societies such as the East Africa Natural History Society, the New England Botanical Club, the International Society for the Study and Conservation of Amphibians; other non profit institutions such as the Bailey-Matthews Shell Museum, the San Diego Natural History Museum, the Peabody Museum of Natural History; and information aggregators such as BioOne have enabled BHL to expand access to more recent, copyrighted publications. Founding institutions like the MBL and Harvard have also scanned their publications like the *Biological Bulletin*, the *Bulletin of the Museum of Comparative Zoology* and *Breviora* with a rolling wall of only one year or none. There are now almost 37,000 titles (more than 26 million pages) in the Biodiversity Heritage Library. The BHL is barely halfway through scanning the total available public domain biodiversity literature. The next larger content loads will come from European and Chinese digitization projects and will greatly enrich the current

BHL corpus.

The multi-institutional nature of the project necessitated the development of collection management and access tools, including the BHL portal to provide access to all of the digitized material regardless of contributing library, a "deduplication" database for monographs to ensure that duplicate copies were not scanned unintentionally, and a serials "bid list" to allow member libraries to indicate serial volumes that they were scanning, again to avoid duplication.

Funding and staff of the BHL-Europe will provide infrastructure development for the BHL. One of the first developments is the creation of redundant repositories based on a system of nodes throughout the world (Figure 2). Some nodes will maintain only local content but there will be multiple complete repositories of BHL data (currently UK and US). Other products to facilitate access and collections are also under development through BHL, BHL-Europe and others.

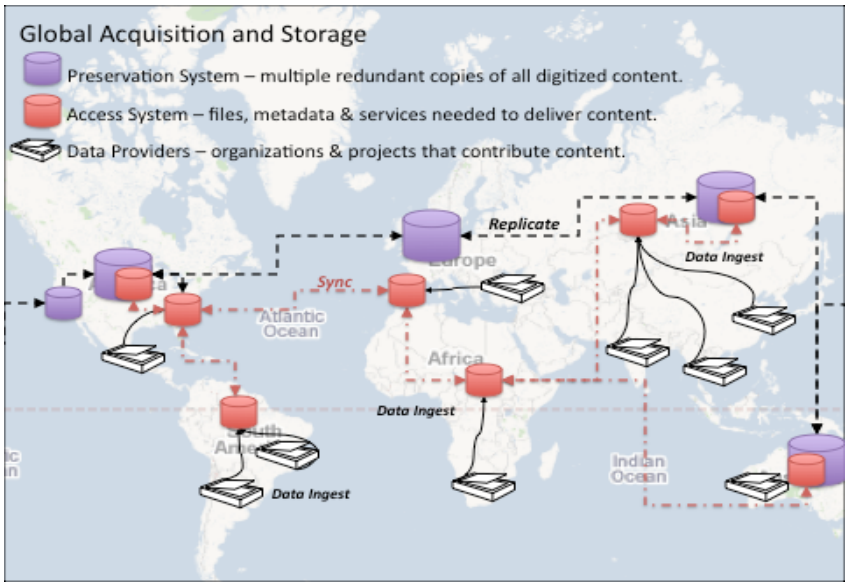

Figure 2: BHL Hub contains entire scope of BHL content & services but tailored for regional-specific or language needs (courtesy of Chris Freeland).

CiteBank is a related repository of scientific citations and community-vetted bibliographies with services linking to other biodiversity projects (http://cite.biodiversitylibrary.org/browse). CiteBank is drupal-based and allows users to upload and share bibliographies containing material related to their specific interests and upload files associated with these bibliographies, including PDFs of the articles and links to the books containing the articles within the BHL portal. Thus, CiteBank is a crowd-sourced, user-dependant service. CiteBank also serves as an interface through which users can network and collaborate, forming groups related to specific interests and subjects. The drupal-based Citebank also presented a solution to the problem of the inflexible interface currently in the BHL portal: the drupal-based interface will become the new portal (Figure 3).

Figure 3: New interface based on Drupal and used for Citebank.

BHL has expanded the expected audience and provides materials to anyone who would otherwise be unable to read and use valuable biodiversity literature. BHL is more than merely a sum of its members' collections: it is vehicle to repatriate information to countries with high biodiversity but few sources of literature. Biodiversity information about many countries resides in North American and Europe. The BHL returns to those countries, via the web portal, vital biodiversity information.

Data integration and referral are critical areas of development for the BHL (Figure 4). BHL links are embedded in EOL and the BHL also refers back to species pages in EOL. Reviewing Google analytics (Figure 5) shows that traffic to the BHL portal is increasing steadily, primarily through referrals from search engines (primarily Google) but also through sites such as EOL and Wikipedia. As part of this process, BHL links were added for major taxonomic works cited in Wikipedia pages about scientists. Referrals from Wikipedia to BHL increased. Adding data from publications such as *Zoological Record* and *TL2* would benefit searching in the BHL. Discussions are underway with the publishers of these tools. Social networking has benefitted the BHL through twitter, Facebook, Library Thing and tagging of unique special collections items.

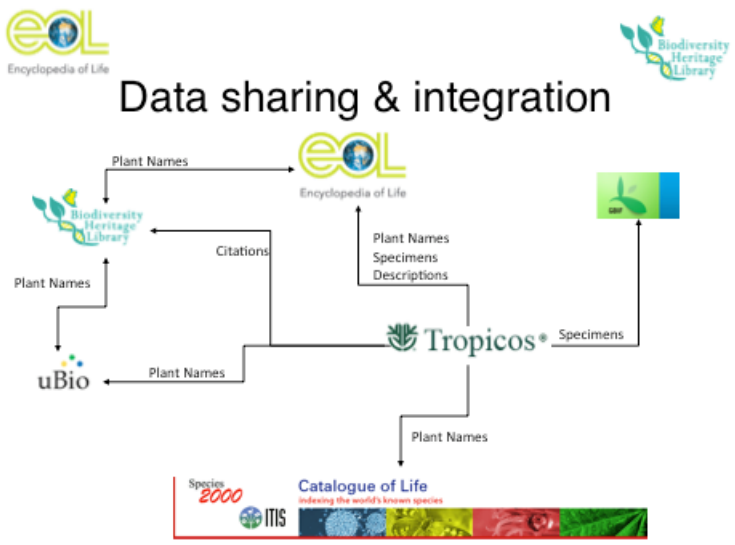

Figure 4: Sophisticated data integration (courtesy of Chris Freeland).

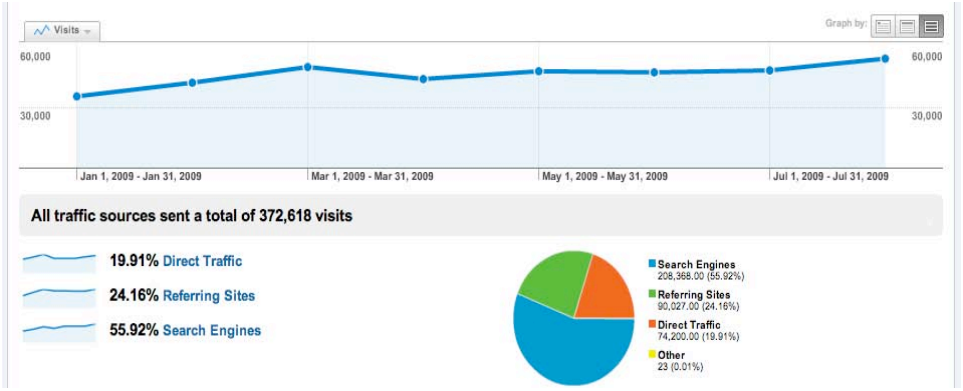

Figure 5: Google Analytics use and referral.

The Harvard University Ernst Mayr Library and Botany Libraries have been the leaders on an IMLS planning grant to find a cost-effective method of bringing mass digitization services to rare and unique materials. This includes bringing together institutions with large special collections and a history of collaborative digitization that can provide solutions for problems like oversized books and foldouts, fragile items, or unusual presentations not currently addressed in the current scanning BHL Program.

The United Nations has declared 2010 as the International Year of Biodiversity to raise awareness about the importance of biodiversity (Convention of Biological Diversity 2009): this is already celebrated in the BHL collaboration (Figure 6) where it

has been demonstrated that multi-library digital collections can provide new ways for users to access materials and provide an opportunity to create a transformative system. The BHL plans to emulate the distribution of IAMSLIC members (Figure 7).

Figure 6: BHL is growing.

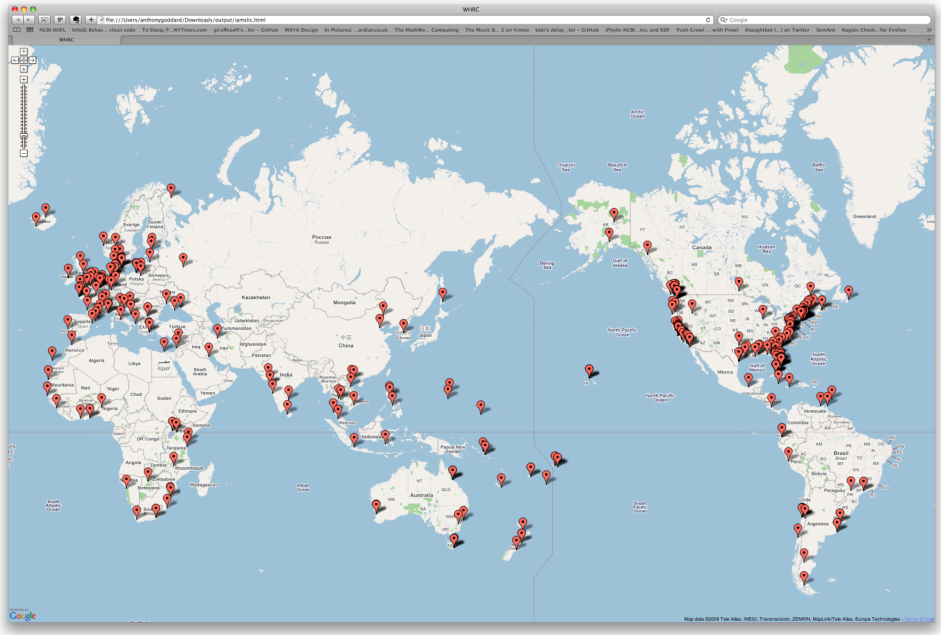

Figure 7: IAMSLIC members, the BHL dream that is quickly becoming reality.

References

Convention on Biological Diversity. 2009. United Nations International Year of Biodiversity. Retrieved December 1, 2009 http://www.cbd.int/2010/welcome/

Gwinn, Nancy and Constance Rinaldo. 2009. The Biodiversity Heritage Library: Sharing biodiversity literature with the world. IFLA Journal, 35(1): 25-34, DOI: 10.1177/0340035208102032

Norton, Cathy. 2008. The Encyclopedia of Life, Biodiversity Heritage Library, Biodiversity Informatics and Beyond Web 2.0. First Monday 13 (8). http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2226/2013

Rinaldo, Constance. 2009. The Biodiversity Heritage Library: Exposing the Taxonomic Literature. Journal of Agricultural & Food Information 10 (3): 259-265, DOI: 10.1080/10496500903014669.

Wilson, E.O. 2007. TED prize wish: Help build the Encyclopedia of Life. Retrieved June 5, 2009 from http://www.ted.com/index.php/talks/e_o_wilson_on_saving_life_on_earth.html