The Biodiversity Heritage Library: Unveiling a World of Knowledge About Life on Earth

Martin R. Kalfatovic[1] [0000-0002-4563-4627], Grace Costantino[2] [0000-0002-5003-7851] and Constance A. Rinaldo[3] [0000-0002-8339-728X]

[1] Smithsonian Libraries, Washington DC 20013, USA
kalfatovicm@si.edu
[2] Smithsonian Libraries, Washington DC 20013, USA
[3] Ernst Mayr Library, Museum of Comparative Zoology, Harvard University, Cambridge MA 02138

Abstract. The Biodiversity Heritage Library (BHL) is an international consortium making research literature openly available to the world as part of a global biodiversity community. Through its extensive network of Members, Affiliates, and Reciprocal Partners, over 56 million pages of biodiversity literature are openly available through the BHL portal. Created in 2006, the BHL was a direct response to the needs of the taxonomic community for access to early literature. The original BHL organizational model, based on United States and United Kingdom partners, provided a template for what is now over 80 global partners.

Now a cornerstone of biodiversity infrastructure, BHL is integral to key databases and data aggregators (e.g. World Registry of Marine Species, Tropicos, Global Biodiversity Information Facility, and the Encyclopedia of Life), and has engaged the research community in tool development and content reuse. Data contained in the Biodiversity Heritage Library (BHL) describes collections held in the world's major museums and botanical garden libraries. Finding those collections data, however, remains a challenge. BHL is actively engaging in incorporating tools and services (including digital object identifiers, full-text-search, and application programming interface) to make finding and linking to collection specimen information better.

Keywords: Knowledge Discovery in Digital Libraries, Research Infrastructures, Digital Libraries, Biodiversity, Life sciences, Biodiversity Heritage Library, BHL.

1 Overview

Operating as a consortium of natural history, botanical gardens, research institutions, and related organizations, the Biodiversity Heritage Library (BHL), founded in 2006, is administered via a Secretariat located at the Smithsonian Libraries in Washington, DC, and operates via global network of staff at participating institutions. BHL collections, spanning the 15th century through yesterday, contain over 56 million pages from over 244,000 volumes, approaching the size of many libraries in typical natural history or botanical collections. Importantly, the BHL data is both human-readable via the BHL portal, but accessible through open APIs that allow use and reuse of BHL data directly by other data systems.

The BHL’s open access collections and services enable scientists to find the information they need to identify, describe, and conserve the world’s species and habitats. The services in the BHL were developed directly in response to the research needs of the scientific community. BHL remains an integral part and responsive to the needs of a growing global biodiversity community.

2 Global Digital Library Partnerships

The BHL is an inaugural and key science data content provider to the Digital Public Library of America in the United States and also to the Europeana, the European Union's portal to museum and library collections. Natural History books and archives provide information that is critical to the study of biodiversity. The species data, ecosystem profiles, distribution maps, illustrations, behavioral, and inter-dependency observations, and geological and climatic records contained in this literature reinforces current scientific research and provides an historical perspective on species abundance, habitat alteration, and human exploration, culture, and discovery. Scientists have long considered lack of access to biodiversity literature a major impediment to the efficiency of scientific research.

3 Open Access and Scientific Impact

Illustrative of the BHL's commitment to open data, the BHL is a charter signatory of the Bouchout Declaration for Open Biodiversity Knowledge Management. BHL’s commitment to open access extends beyond making scanned pages available through BHL. Content is available via Internet Archive, the Digital Public Library of America, and Europeana, and BHL’s suite of APIs brings BHL data directly to users. BHL metadata is licensed as CC0 to allow for the widest dissemination of the data.

As noted in the Bouchout Declaration: “Collaborative Open Biodiversity Knowledge Management can bring together the achievements of many independent biodiversity projects, yet will allow them to retain their identity and missions. The resulting virtual pool of information will allow new services to emerge for everyone who relies on information about life on Earth. Awareness of, access to, preservation, and curation of information will be enhanced by a shared and seamless network of infrastructures” [9].

As noted by Hobern, et al., BHL is an important part of the global biodiversity ecosystem: “As an example of an existing collaboration, which might serve as a proof-of-concept project and which would benefit from increased exposure and openness, Catalogue of Life, GBIF, Encyclopedia of Life, Barcode of Life Data Systems and Biodiversity Heritage Library are currently working to develop a new collaborative model for building a shared taxonomic framework, under the project name, Catalogue of Life Plus (CoL+)" [11].

4 Technology

The BHL portal is located at the Smithsonian and runs under a .Net environment. BHL offers various web APIs, downloads, and other services for export and download of BHL data and content. BHL partners with the Internet Archive for file staging and storage. Additionally, BHL has developed software service, MACAW that simplifies ingest of partner content into BHL. MACAW is an open source tool with all code available on GitHub [16].

Looking forward, the BHL is in the planning stages of the next iteration of the BHL portal, dubbed BHL EVO, that will provide a number of enhancements and redesign elements, including the following:

- **Platform Transformation.** A total revamping of the BHL back-end will move the platform to the latest proven technologies. Funding will allow for the contracting of two developers/programmers to rewrite the underlying BHL code-base and transition the hardware to new open source technologies. Modularizing the BHL code will enable more rapid development in the future and create opportunities to co-develop with other major digital libraries such as the HathiTrust, Digital Public Library of America, etc.
- **Metadata Model Revision.** The fundamental recommendation is to create a concept of derivative digital objects (DDOs) that are first-class citizens, and to make it easy to create both individual DDOs and new classes of them. DDOs might eventually represent such diverse objects as ebooks, articles (including born-digital articles that are not part of an Item), chapters, snippets, annotations, plates, lists, tables, interesting flies squashed between two pages, and more. Once a DDO has been created, it must be easy to enrich it and to manage links to it and from it, and to have them form part of larger wholes within or outside of the BHL.
- **Semantic and Linked Data.** Enhancing existing BHL data and metadata through linked data and sematic web technologies will allow for more machine-to-machine use of deep BHL content. This data, which includes key facets such as habitat, diet, and author, and biomorphic data, will be exposed for researchers to integrate it into their own tools and publications. This data enhancement will created expanded links to key biodiversity resources such as the existing links to Tropicos, GBIF, and the Encyclopedia of Life.
- **Enhanced Platform.** The BHL platform excels at delivery of books. Funding will allow for improvements to the platform to enhance current field book content discovery and display; provide better support for Arabic and Asian-language texts; provide the capabilities to delivery new types of content including maps, visual resources, and related biodiversity content held in partner collections. Additionally,

this will include integrated search and discovery of the rich illustrations and visual treasures included in the BHL corpus and deliver this content to new communities. Additional human interface improvements will include multilingual support and personal customization of the interface by users and the ability to save searches.

- **Dimensions of Biodiversity Literature and Gaps in BHL.** Finding a number to describe the quantity of biodiversity literature available is elusive. As BHL evolves into BHL Version 2 and content continues to become more robust, it is important to identify what is missing from the BHL collections. Identifying gaps will provide targets for prioritization and funding. Users have requested the incorporation of more in-copyright content providing another option to pursue for adding collections.

5 Conclusion

The BHL has been a successful model for digital library development and collaboration for a variety of reasons. Chief among these has been the BHL focus on a specific use case, namely, providing access to a core constituency of taxonomists. Expanding beyond this core group in carefully planned stages (e.g. those interested in the visual images within BHL) and partner growth (e.g. the global growth via targeted institutions) has broadened both the participant and user base. Likewise, the integration of BHL into core functions of its participating institutions has put BHL on the path to financial sustainability. BHL looks to the future with planned and sustainable growth, enriched content, and appropriate services.

References

1. Biodiversity Heritage Library joins GBIF as associate, https://www.gbif.org/news/82358/biodiversity-heritage-library-joins-gbif-as-associate, last accessed 2019/04/24
2. The Bouchout Declaration for Open Biodiversity Knowledge Management, http://www.bouchoutdeclaration.org/declaration/, last accessed 2019/04/23.
3. Rinaldo C., Ford L., deVeer J.: Museum, Library and Archives Partnership: Leveraging Digitized Data from Historical Sources. Biodiversity Information Science and Standards 2(e25920) (2018). https://doi.org/10.3897/biss.2.25920
4. Hobern D., Baptiste B., Copas K., Guralnick R., Hahn A., van Huis E., Kim E., McGeoch M., Naicker I., Navarro L., Noesgaard D., Price M., Rodrigues A., Schigel D., Sheffield C., Wieczorek J.: Connecting data and expertise: a new alliance for biodiversity knowledge. Biodiversity Data Journal 7 (e33679) (2019). https://doi.org/10.3897/BDJ.7.e33679
5. Reaka-Kudla, ML., Wilson, DE., Wilson, EO., eds. Biodiversity II, Understanding and Protecting Our Biological Resources. Joseph Henry Press, Washington, DC (1970).
6. Biodiversity Heritage Library Collection Development Policy (2015), https://s.si.edu/BHLcollectiondevelopment, last accessed 2019/04/24.
7. Parilla, L., Blase, J.: The Value of Flexibility on Long-term Value of Grant Funded Projects. D-Lib Magazine 21(9). DOI:10.1045/september2015-parilla
8. macaw-book-metadata-tool, https://github.com/gbhl/macaw-book-metadata-tool, last accessed 2019/04/24.