

**OPEN ACCESS TO LEGACY TAXONOMIC LITERATURE: INDEX
ANIMALIUM
&
THE BIODIVERSITY HERITAGE LIBRARY: OPEN ACCESS TO LEGACY
LITERATURE**

Martin R. Kalfatovic

Head New Media Office and Preservation Services Department
Smithsonian Institution Libraries
P.O. Box 3712
Natural History Building
Washington, D.C. 20013-3712

Courtney Anne Shaw, PhD

Vertebrate Zoology Librarian
Smithsonian Institution Libraries
P.O. Box 3712
Natural History Building
Washington, D.C. 20013-3712

Suzanne C. Pilsk

Librarian Metadata Specialist, Technical Services Division.
Smithsonian Institution Libraries
P.O. Box 3712
Natural History Building
Washington, D.C. 20013-3712

Abstract : After reviewing taxonomic literature, the principles of nomenclature, and introducing Sherborn's *Index Animalium*, the speakers talk about going beyond page scanning to data parsing and data mining, thus being able to connect users to the literature they need. The Biodiversity Heritage Library, a project to digitize the monographic and serial taxonomic literature is then discussed.

Keywords: Charles Sherborn, Index Animalium, species, taxonomic nomenclature, digitization, bibliographic references, OCLC WorldCat, Smithsonian Institution Libraries, National Museum of Natural History, Marine Biological Laboratory, UBio, Open Content Alliance, Biodiversity Heritage Library

I. Index Animalium Digitization Project

In any well-appointed Natural History Library there should be found every book and every edition of every book dealing in the remotest way with the subjects concerned. ... Moreover for accurate work it is necessary for the student to verify every reference he may find; it is not enough to copy from a previous author; he must verify each reference itself from the original.¹

Charles Davies Sherborn, a noted taxonomist, clearly states the relationship between biological nomenclature and the need to reference published works. Sherborn's *Index Animalium* was, at its time of publication, the definitive index to animal names. This important link between working scientists and the reference materials stored in libraries is the reason for the current work being done to provide greater access to these printed texts. Through the digitizing of Sherborn's *Index Animalium* and the further step of development of a world wide accessible Biodiversity Heritage Library, the Smithsonian is participating in the globalization of these important texts.

A. Binomial nomenclature

Binomial nomenclature is the standard convention used for naming species. As the word 'binomial' means, the scientific name of a species is formed by the combination of two terms: the genus name and the species epithet or descriptor. The naming of a species is done by an "author," the person who first publishes the name. The species descriptor should be an adjective that differentiates a species from other members of a genus. The genus name and species descriptor are usually derived from Latin but more modern naming conventions have developed to be more "Latin-like." Geographic features (cities, mountains, rivers, etc.) are used to form the description or some are named after prominent people. For example, F. Christian Thompson, a USDA entomologist, described and authored a flower fly, which he named after Bill Gates: *Eristalis gatesi*. (Bill's fly is only found in the high mountain cloud forests of Costa Rica).

Established rules exist for the proper naming in the various fields of study. But in every field, zoology to botany, the name is considered established once it has appeared in a published document. Rules established in the *International Code of Zoological Nomenclature*² state that the documents must be distributed in at least five major, publicly accessible libraries. This creates a logical, clear, and important tie between the library community's stewardship of printed, published documentation to the scientific community in the taxonomic field. Taxonomic literature never goes out of "style". It remains necessary and even at times more important as it gets older.

¹Charles Davies Sherborn, Epilogue to *Index Animalium*, March 1922.

² <http://www.iczn.org/iczn>

B. The *Index*

Charles Sherborn was a cataloger at heart. He examined relevant text looking for names, creating a hand written card index that was useful as soon as he started. Published between 1902 and 1933 in two section comprised of 33 parts, the *Index Animalium* covers a range of species and genera names giving the exact location of the name in the published text. Smithsonian Institution Libraries has completely digitized Sherborn's *Index Animalium* providing scanned images of the pages and searchable database of the information. Currently, staff at the Smithsonian is in the process of deciphering the abbreviations used by Sherborn and making the logical and explicit connection between the references to bibliographic records for each text cited.

A typical Sherborn entry:

albimanus Delphinus, T. R. Peale in Wilkes, Expl. Exped. VIII. 1848, 33

Sherborn's index differs from traditional citations by giving the species name first. The dolphin *Delphinus albimanus* was authored by Titian Ramsey Peale. Peale named it for the first time in the eighth volume of Charles Wilkes' multivolume publication *United States Exploring Expedition*. Volume 8 was published in 1848 and Peale's description and name for this species is on page 33.

The Atherton Seidell Endowment Fund of the Smithsonian Institution supported the Smithsonian Libraries in converting the OCR text to 99.997% accuracy. SIL worked in collaboration with a team from the Marine Biological Laboratory at Woods Hole to take the re-keyed text and parse the data³. David Remsen and Patrick Leary had worked on a similar text, Neave's *Nomenclator Zoologicus*⁴. Remsen and Leary used a combination of PHP routines and regular expressions to create a database fielding the names, the authors, the publications and other information from Sherborn's text.

C. Bibliographic connections

The bibliographies of *Index Animalium* hold particular interest. All the texts consulted by Sherborn, monographs and journals, are listed. Sherborn used non standard abbreviations for the titles and publishers making it difficult to identify which text he examined. The goal of the bibliography abbreviation project is to create a full citation for each of the text mentioned in the *Index Animalium* bibliographies, and to provide a connection to a full bibliographic record for the title. Once identified, the title and author are searched to find linkable bibliographic descriptions of the text in SIRIS⁵, the Smithsonian's online

³SIL is posting images of the pages as well as pictures of pages because these systematists and taxonomists care to see exactly what was published noting errors of Sherborn and errors of the publishers.

⁴<http://www.ubio.org/NomenclatorZoologicus/>

⁵Smithsonian Institution Research Information System. <http://www.siris.si.edu/>

catalog, and OCLC's WorldCat. The final web product will link the users from the citations and bibliography to the texts in the Smithsonian collection and to WorldCat's worldwide library catalog.

Screen shot of Bibliography Abbreviation Project:

Original Text	Full Title	Authorized Name	OCLC	SIRIS
Richardson, John. Fauna Bor.-Amer. 4 pts. Lond. 1829-37. [Quadr. June 1829 ; Birds, Feb. 1831 ; Fish, 1836 ; Ins. 1837.],,Yes	Fauna boreali-americana; or, The zoology of the northern parts of British America: containing descriptions of the objects of natural history collected on the late northern land expedition, under command of Captain Sir John Franklin.	Richardson, John, Sir, 1787-1865.	4055433	185713
Riedel, W. Die Grasmücken.... 8vo. Nördl. 1833. [B. M., no n. spp.],,	Die Grasmücken und Nachtigallen in Europa, oder, Vollständige Naturgeschichte dieser vorzüglichsten Singvögel :nebst Zaunkönig und Goldhähnchen : mit besonderer Rücksicht auf Fang, Zählung, Pflege, Wartung, Nutzen und Vergnügen : ein unentbehrliches Han	Riedel, Wilhelm, Pfarrer in Pfuhl	19469019	364051
Risso, A. Ichthyol. de Nice. 8vo. Paris, 1810.,,	Ichthyologie de Nice, ou Histoire Naturelle des Poissons du department des Alpes Maritimes.	Risso, A. (Antoine), 1777-1845	19469044	364052

II. Biodiversity Heritage Library Project

*Yet another physical difficulty is the task of assembling the library and indexes which will enable the student to work under proper conditions....The beginner must now be prepared to spend liberally, or else must establish himself in an institution where a large library exists; if he work by himself with only a few books, he will have to confine himself to a very narrow specialty indeed.*⁶

James M. Aldrich, a Smithsonian entomologist quoted above, had a dream to bring together all the taxonomic publications into one large library. The Convention of Biological Diversity held in Darwin Australia, February 1998 noted in its Darwin Declaration of Life⁷ that the existing information held in the literature and by current experts should be made available electronically. Modernizing this dream, the Biodiversity Heritage Library (BHL) was formed to create a digitized, virtual collection of taxonomic literature. The *Index Animalium* will provide a basis of texts to be used – ideally, every publication listed in *Index Animalium* will be digitized accessible through the BHL.

A. Example of successful digitization

The Smithsonian has found that digitizing texts and making them available over the World Wide Web has been incredibly beneficial to researchers world wide. It provides a “repatriation” of information to areas of the world that do not have access to those legacy texts describing their own biological diverse ecosystems. *Biologia Centrali Americana*, a multivolume compendium of the biodiversity of Mexico and Central America at the turn of the 19th century, has only few complete copies in North America, fewer in Europe, and only two in Central America. The two in Central America are housed in Smithsonian facilities. This limited distribution has been solved by providing the digitized text.⁸ This allows researchers in the field to access data, though possibly through very slow connections, directly instead of requiring very long distant traveling to the few institutions that hold the hard copies of these types of materials.

B. History

The idea of providing as much data as possible to scientists where ever they maybe doing research is not new. In 2003, a meeting took place in Telluride, Colorado to discuss the potential of creating Edward O. Wilson’s concept of the “Encyclopedia of Life:”

Imagine an electronic page for each species of organism on Earth, available everywhere by single access on command. The page

⁶“The Limitations of Taxonomy” by J.M. Aldrich, *Science*, April 22, 1927, vol. LXV, no. 1686, p.381.

⁷ <http://www.biodiv.org/programmes/cross-cutting/taxonomy/darwin-declaration.asp>

⁸ e-Biologia Centrali- Americana <http://www.sil.si.edu/digitalcollections/bca/>

contains the scientific name of the species, a pictorial or genomic presentation of the primary type specimen on which its name is based, and a summary of its diagnostic traits.⁹

In February of 2005, a meeting took place in London: “Library and Laboratory: the Marriage of Research, Data and Taxonomic Literature.” From that meeting, natural history librarians took the idea of a combined digital library to Washington in May of 2005. Ground work for the Biodiversity Heritage Library grew from that and has continued with an organizational and technical meeting again in Washington the summer of 2006. The goal of having information linked providing seamless access for users to look up species, verify information in texts, link to updated species information, references to historic usage of names, accurate images of species, and even geographic distributions is not as far off as it seems.

C. BHL Membership

BHL consists of five large natural history museums: American Museum of Natural History (New York), National Museum of Natural History, Smithsonian (Washington, D.C), Natural History Museum (London), and the Field Museum (Chicago). Three major botanical gardens: Missouri Botanical Garden, New York Botanical Garden, and the Royal Botanic Garden, Kew. Botany Libraries and the Ernst Meyer Library of the Museum of Comparative Zoology at Harvard University are also members. The informatics member is the Marine Biological Laboratory / Woods Hole Oceanographic Institution Library (Massachusetts). Partnering for digitization is the Internet Archive (San Francisco).

BHL is also a part of the Open Content Alliance¹⁰. A major mandate of the BHL and OCA’s digitization efforts is to provide open access. The plan is to include all material that is out of copyright and having the “opt in” model (not the Google “opt out” model) for publishers of currently held copyrighted materials.

D. Taxonomic Intelligence

The uBio initiative at the Marine Biological Laboratory Library is an international effort to create and utilize a comprehensive and collaborative catalog of known names of all living (and once-living) organisms.¹¹ UBio’s algorithm for harvesting taxonomic binomial names from OCR text and adding and comparing it to their growing Name Bank of species names allows for taxonomic identification. This allows for synonym reconciliation – and even has vernacular tools, Roman and Non –Roman script

⁹E.O. Wilson, “The Encyclopedia of Life”. <http://www.all-species.org/fall/references/EncyclopediaofLife.pdf>

¹⁰<http://www.oca.org>

¹¹<http://www.ubio.org/>

capabilities – and then connects the information to other taxonomic sources such as ITIS or Species 2000.

Biodiversity Heritage Library - Process

Processing: http://names.ubio.org/bulletin/39088006090120_divu.xml

Bulletin of the Bureau of Fisheries Vol. XXXI Part II

Pages examined:	333/333 (100%)	
Total Scientific Names:	7376	
Distinct Scientific Names:	3497	
Total in Namebank:	2488 (71%)	View
Additional related names		
Scientific Synonyms:	4555	
Vernacular Synonyms:	10501	
Total Names available for search		
without Taxonomic Intelligence:	3497	
with Taxonomic Intelligence:	17544 (+502%)	View

Note that a percentage of names within the volumes are currently not cataloged within NameBanks 6 million records. This means that they also are not a component of any of the global species datasets we index including Species 2000, ITIS, NCBI, etc. This is important because we can group and present these uncataloged names to these aggregators for taxonomic vetting.

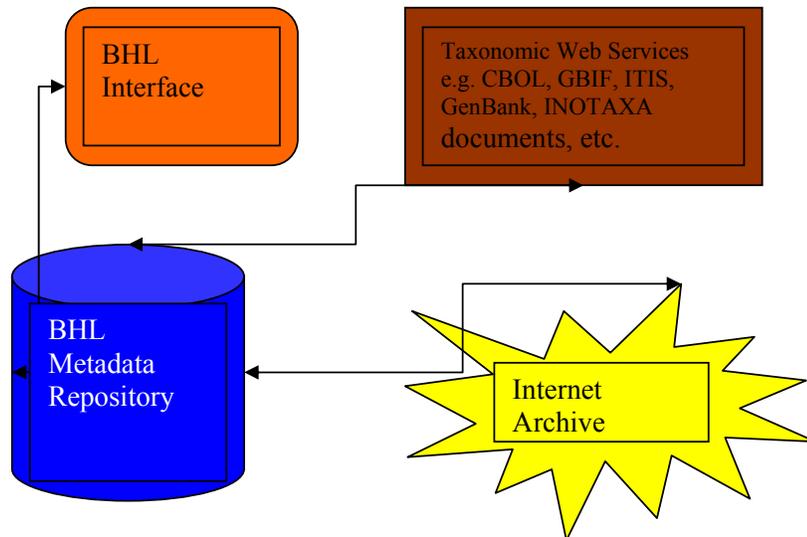
Best Fit Classification: **Species 2000**

 **Original Names Mapped:** 1697 [Browse](#) 

NameBank Additions based on Synonyms: 5962

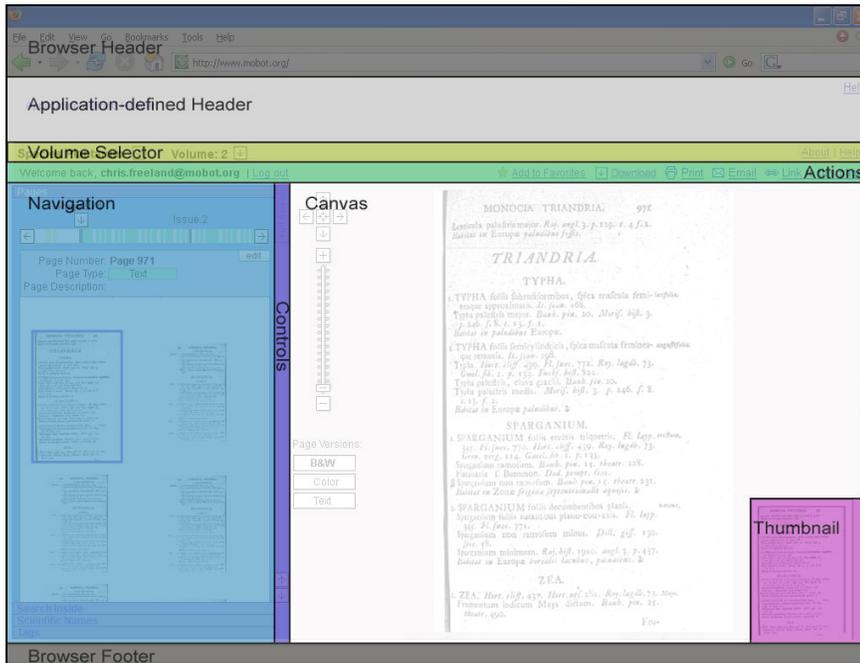
In the above example, a sample text is put through the tool; working against the NameBank list of over 9.5 million names, nearly 1,000 new names were located and over 17,000 new access points (valid names, synonyms, vernacular terms, common misspellings, etc.) are now available to the researcher. The displays also allow browsing of texts by taxonomic tree sets.

E. Future Goals of BHL



A very simplified schematic of BHL indicates that there may be many ways to access the information. BHL will have an interface or portal, but the data will be available for other services to use the data exposed by the BHL. The metadata repository will hold the title level description, plus some granular level identification needed for taxonomic citation (volumes, issues, etc.). The metadata repository can point to the files stored at the Internet Archive. These files will include the scanned images, OCR text and other related files.

Chris Freeland, of the Missouri Botanical Garden, has designed a prototype of what the BHL interface might include. The wireframe holds scanned images, navigational menus, and functions for ways of capturing, storing and printing out materials.



The short term goals for the BHL project are to continue the analysis of the metadata from the member institutions. Development of scanning workflow and plans are being discussed as are the locations of potential high production scanning stations.

Long term goals for BHL include fostering relationships with scholarly publishers of current taxonomic journals; working on the metadata needed at the levels of citation in taxonomic texts; integrating with proposed the Encyclopedia of Life; work with the international biodiversity organization, GBIF; and coordinate efforts with global taxonomic databases such as Consortium for the Barcode of Life and the National Center for Biotechnology Information's GenBank.

REFERENCES

Index Animalium / Charles Davies Sherborn : an electronic edition by Smithsonian Institution Libraries. <http://www.sil.si.edu/digitalcollections/indexanimalium/>

Biodiversity Heritage Library. <http://www.bhl.si.edu/>

Universal Biological Indexer and Organizer (UBio). <http://www.ubio.org/>

<http://www.sil.si.edu/staff/2006IAMSLIC/PortlandIAMSLICPresentationfinal.pdf>