Journal of Library Metadata, 10:136–155, 2010

ISSN: 1938-6389 print / 1937-5034 online
DOI: 10.1080/19386389.2010.506400

Routledge
Taylor & Francis Group

The Biodiversity Heritage Library: Advancing Metadata Practices in a Collaborative Digital Library

SUZANNE C. PILSK
Smithsonian Institution Libraries, Washington, DC, USA

MATTHEW A. PERSON
MBLWHOI Library, Marine Biological Laboratory, Woods Hole, Massachusetts, USA

JOSEPH M. DEVEER
Ernst Mayr Library, Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts, USA

JOHN F. FURFEY
MBLWHOI Library, Marine Biological Laboratory, Woods Hole, Massachusetts, USA

MARTIN R. KALFATOVIC
Smithsonian Institution Libraries, Washington, DC, USA

The Biodiversity Heritage Library is an open access digital library of taxonomic literature, forming a single point of access to this collection for use by a worldwide audience of professional taxonomists, as well as "citizen scientists." A successful mass-scanning digitization program, one that creates functional and findable digital objects, requires thoughtful metadata work flow that parallels the work flow of the physical items from shelf to scanner. This article examines the needs of users of taxonomic literature, specifically in relation to the transformation of traditional library material to digital form. It details the issues that arise in determining scanning priorities, avoiding duplication of scanning across the founding 12 natural history and botanical garden library collections, and the problems related to the complexity of serials, monographs, and series. Highlighted are the tools, procedures, and methodology for addressing the details of a mass-scanning operation. Specifically, keeping a steady flow of material, creation of page level metadata, and building services on top of data and metadata that meet the needs of the targeted

Address correspondence to Suzanne C. Pilsk, Smithsonian Institution Libraries, P.O. Box 37012, MRC 154, Washington, DC 20013-7012, USA. E-mail: pilsks@si.edu

communities. The replication of the BHL model across a number of related projects in China, Brazil, and Australia are documented as evidence of the success of the BHL mass-scanning project plan.

KEYWORDS Biodiversity Heritage Library, taxonomic literature, digital libraries, digitization projects, digitization workflow, mass-scanning projects, collaboration, natural history libraries

"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" (Sherborn, 1932).

BIRTH OF THE BIODIVERSITY HERITAGE LIBRARY

The early 2000s saw the enthusiastic embracement of developing digital Web technologies across various fields of study and methods of research. New approaches to traditional work were attracting research institutions, natural history museums, taxonomists, and libraries. This was the beginning of bridging across silos of information and closed communities of practice to create integrated communities of knowledge. Natural history libraries saw this as an opportunity to explore how to better support the needs of taxonomists, nomenclaturists, and the general species-identifying community. In 2005 a meeting was held at the Natural History Museum in London, referred to by participants as LibLab: Library and Laboratory; the Marriage of Research, Data and Taxonomic Literature. This meeting helped to elucidate a pending "perfect storm," defined by the confluence of reasonable scanning costs, significant scannable library collections, and a geographically dispersed and demanding user community. The outcome was clear; there was a need to move toward the implementation of a global digital library project.

At the time it appeared to be a novel idea, but it was clear that scientists around the world would use a digital library of taxonomic literature. Borrowing a term from another scientific field, the *half-life* of taxonomic literature is longer than that of any other scientific arena (Moritz, 2005). In most other sciences, the immediate need is for current publications revealing the most recent discoveries and the most salient laboratory data. In taxonomy, it is the historical literature that is critical for the identifying and naming of species. The state of the art of scanning and serving literature via the Internet was mature enough to handle the needs of this community. Old, rare, and required literature could be provided electronically wherever and whenever the researcher needed it.

To move forward with the successes discussed at LibLab, a partnering group formed naturally from participants and produced a memorandum of understanding establishing the Biodiversity Heritage Library (BHL). Initial library members included the American Museum of Natural History (New

York, NY), Harvard University Botany Libraries (Cambridge, MA); the Ernst Mayr Library of the Museum of Comparative Zoology, Harvard University (Cambridge, MA); MBLWHOI Library of the Marine Biological Laboratory and Woods Hole Oceanographic Institution, (Woods Hole, MA); Missouri Botanical Garden (St. Louis, MO); the Natural History Museum (London); New York Botanical Garden, LuEsther T. Mertz Library (New York, NY); the Royal Botanic Gardens, Kew (London); and Smithsonian Institution Libraries (Washington, D.C.). In May 2009, two additional members, the Academy of Natural Sciences (Philadelphia, PA), and the California Academy of Sciences (San Francisco, CA) joined the consortium. Funding was made available through a grant from the MacArthur Foundation (via the Encyclopedia of Life http://eol.org/) to begin digitization with the Internet Archive as the BHL scanning partner.

The mass-scanning project was launched via easy decisions stemming from a clear focus on biodiversity literature, combined with a realization of the need to work out challenging workflows. Librarians at the partner libraries clearly defined the scope of the materials to be scanned: anything out of copyright that was critical to the work of their clientele. The challenges that emerged were identifying the specific titles, avoiding redundant work, and, if possible, avoiding duplication of scanning. To achieve economies of scale, the three mass-scanning centers (located in New York City, Washington, DC, and Boston) established by Internet Archive demanded a sizable quantity of material. The "mass" scanning approach insisted on an effective and efficient work flow to process the quantity required by the Internet Archive "beasts." "Feeding the beast" became a catch phrase to each BHL member library that described the need to keep the material flowing to the scanning centers. The operation needed to run continuously and the quality of the finished digital product had to be acceptable.

As much a sociology project as a scanning project, BHL gave rise to a multi-institutional team of international librarians that grew organically from the partnering libraries in an extremely positive and collaborative atmosphere. In other words, a major asset has been the collegiality and agreement among team members on the fundamental direction of the project. The librarians on the frontline have worked out effective ways to communicate, share expertise, and lend helping hands to one another for the good of the project and to ensure a successful initiative. This paper is indicative of the cooperative spirit underlying the BHL. In fact, the American Library Association is recognizing the collaborative nature of the project by awarding the 2010 ALCTS Outstanding Collaboration Citation to BHL.

TAXONOMISTS AND LIBRARIES

Natural history libraries have historically been in a unique position to support the work of taxonomic nomenclature due to the requirements of the

field. In all areas of systematic taxonomy, scientists identifying and naming species follow specific rules and guidelines based on a system created by the Father of Taxonomy, Carl von Linnè, a.k.a Linnaeus. During a recent broadcast of the popular American television show *Jeopardy!* a final question referred to the mnemonic that most students use to remember the Linnaeus' outline: "Kings Play Chess On Fine Grain Sands" equals Kingdom, Phylum, Class, Order, Family, Genus, Species. The Linnaen system of classification is the art of naming species resulting in a system that is orderly, not overly redundant, attempts to avoid duplications and synonyms for the same species, and can be communicated across disciplines, languages, and oceans. Each taxon community has its own rules for its specific area of expertise (e.g., botanists and zoologists each have unique sets of rules, and specific taxon subgroups will choose a specific classification scheme). A major element in all the rules agreed upon by the nomenclaturists who are identifying, naming, or revising species—whether a bug or a sprout, endoskeletal, exoskeletal, or invertebrate—currently involves the printed published literature.

The preamble of the International Commission on Zoological Nomenclature's International Code[1] states: "The objects of the Code are to promote stability and universality in the scientific names of animals and to ensure that the name of each taxon is unique and distinct." The code specifies that "the name or nomenclatural act must have been published" (Article 11.1) and the publication must be freely available to the public. The rules continue with specifically mentioning libraries as holders of the published records.[2]

Natural history libraries are critical for preserving and maintaining taxonomic information. Libraries are mandated by these rules to store the publications that name species as they are discovered. As curators of these library collections, librarians are obligated to have freely available copies for all scientists doing research on species to review and discover what has been found, documented, and named. In general, larger and well-funded natural history libraries have been hosting natural history publications since the 1700s. Since that time, researchers have traveled to these collections or borrowed from them to conduct their work. Systematic taxonomy publications require species citations from the published works. In other words, the published literature is critical to discovery, revising, and naming of life.

Usually found in the form of a "Linnaean binomial," the Latin-modeled name is used to name the organism. A complete taxonomic citation includes the original scientist's name (the "author" of the species) and date of discovery. Typical naming is familiar to most: "*Homo sapiens*" (Linnaeus, 1758) translates to genus *Homo* and species *sapiens*, as named by Linnaeus in his published work of 1758. The 1758 edition of Linnaeus' major work *Systema Naturae*, was the first complete edition and is generally considered to be the starting point for taxonomic nomenclature.[3]

PROBLEMS IN SOLVING REAL NEEDS OF TAXONOMISTS

Over the centuries, scientific nomenclature specialists have been forming species citations using abbreviations and notes that accommodate their special fields of study. These practices developed in an isolated manner specific to discipline (or subdiscipline) and independent of related species projects or the expertise of librarians and information professionals. Concurrently, librarians developing and implementing metadata policies and procedures never entered into conversation with taxonomists. Librarians did little to discover how taxonomic citations are formed and what the actual access needs of the scientists were. As a result, each group adopted its own way of processing information, and constant translation between the two worlds was necessary. At best, a clumsy, but acceptable, disconnected reliance existed between these two fields.

On the scientific researcher side of things, species citations define an "author" of a species as the scientist that identified and named the living thing, not the authors of the book that holds the citation. Abbreviations are used throughout taxonomic citations. The example of *Homo sapiens* was given above. Another is the taxonomic classification of the standard goldfish: "*Carassius auratus* (Linnaeus, 1758)." Another example of a typical descriptive citation is that for the dolphin found on page 133 of the *Catalogue of the specimens of Mammalia in the collection of the British Museum,* published in 1850 by the British Museum Department of Zoology:

> *Delphinus albimanus,* J. Peale, U.S. Exp. Exped. 33 (t./f. Lined.). Snout, head, back, tail and dorsal fin blue-black; belly and pectoral fin white; sides pale tawny; eyes small, brown, and surrounded with a black ring, which joins the black of the snout; body between the dorsal fin and tail very much compressed (Gray, 1850).

In the above taxonomic citation, J. Peale is the person who named or authored the scientific binomial *Delphinus albimanus*, publishing his work in the *U.S. Exploration Expedition*. The above citations raise a series of questions that impact identification and understandability and the ability to resolve to the proper book, volume, and page. Obvious problems arise from the lack of standard bibliographic description as expected by librarians, including the main entry or access point.

The abbreviations of the titles, a single date, and an author that might not be considered even traceable in traditional library cataloging of a monographic series frequently lead these citations to dead ends in the traditional intergrated library system. A trained librarian can read and translate, yet a computer resolving to a digitized book fails. Traditional cataloging lacks the access points of the commonly used bibliographic short or brief title. The ISBD punctuation does not translate to the standards used by the taxonomist.

It seems that the library community is unaware of the frustration some scientists experience in trying to decipher library metadata. As the BHL project began initial planning, informal interviews with potential users were conducted. A surprising conversation with a botanist was revealing. Although she fully supports libraries and her home institution research library in particular, she relies primarily upon basic botanical reference tools to find what she really needs when doing her research. She was sympathetic to the library's need to place a physical object in one place on the shelf and have it retrievable by basic access points, but the amount of detail she requires is not contained in those access points. Furthermore, the data that is exposed is in a format that needs cross-walking to botanical metadata.

Librarians not familiar with a given discipline were surprised to learn the number of abbreviations and amount of assumptions made in the citations that the specialist understands and translates quickly. This makes for difficult computer-to-computer resolving within this new digital world. The mass scanning of the literature solves the issues related to getting the material out to the community, but the overwhelming challenge of translating between century-old systems of notations is not as forthcoming.

MASS SCANNING WORK FLOW AND METADATA ACHIEVEMENTS

Funding requirements created deadlines and specific deliverables expected of the BHL. The funds from the *Encyclopedia of Life*[4] ($50 million, 10-year landmark project to develop one Web page per known species) necessitated the need to start scanning swiftly during the Fall of 2007. Initially the BHL partners decided to forego a formal analysis of our subject-specific collections data. However, the OCLC Worldcat Collection Analysis tool was used for 1 year to help the individual administrative staffs of the BHL partners make subject-specific decisions, support granting requests, and gain a coherent overview of collection strengths. The results did not, in a practical sense, get the books to the scanners.

The mass-scanning assembly-line–like work flow initiated a need to solve some basic bibliographic communication issues that relied on the metadata used by the partnering libraries. Some important lessons that developed out of the Biodiversity Heritage Library (BHL) project were in an attempt to combine metadata elements to drive the scanning and describe the resulting titles, volumes, and pages accurately. BHL library staff members met in person, on the telephone, via email, and through wiki entries, discussing these issues to resolution. They pooled their strengths and expertise and developed critically helpful tools (detailed below), met with Internet Archive programming staff, and put forth workable solutions to begin the initial and critically successful scanning that took place during the first 2 years of the project (2007–2009).

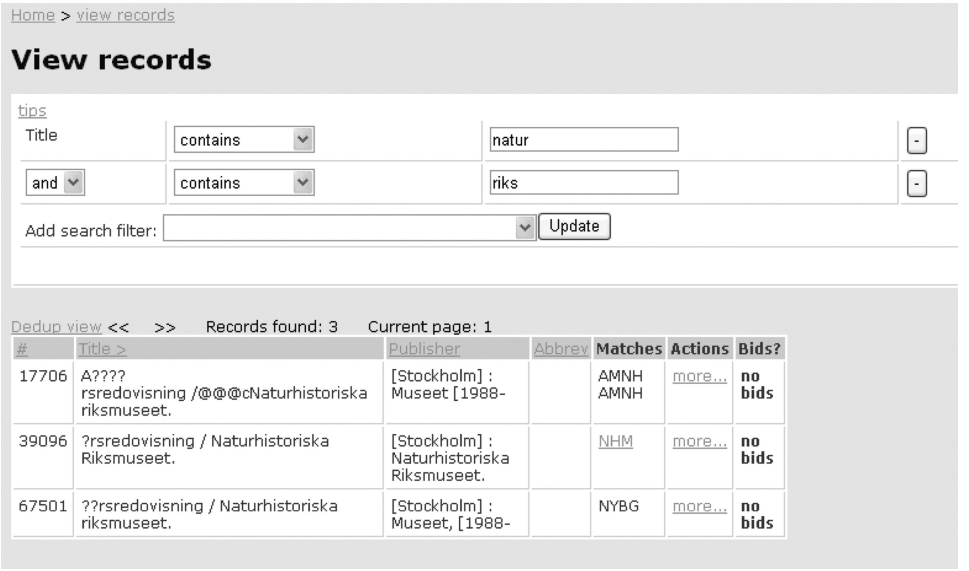

FIGURE 1 View of serials bid list from 2007 (Diacritic display issues have since been resolved).

SERIAL SELECTIONS

The Natural History Museum in London stepped up to the cricket batting box and worked on a program using CakePHP to construct a tool that would enable development of an effective serials scanning work flow.[5] The Serials Bidding List tool (initially known as the Serials Mashup and Union List) was constructed (see Figure 1), and all BHL partners populated the tool by providing MARC dumps of their serials records and holding statements. Out of an initial file of 119,377 records, matching was performed using OCLC numbers, ISSNs, and title (245), out of which 70,764 unique titles were identified. This union list of serials was then viewable and editable by all BHL members.

The Serial Bidding List gives clear indication of the library claiming a title they intend to scan, as well as volumes and years. The tool and the agreed upon procedures allow for subsequent editing of the bid to reflect which titles, volumes, and years actually did get scanned. It also allows other libraries to bid on the same title, scanning volumes that were not completed by the initial partner. Manual deduping and merging to consolidate multiple institution records for single titles is performed on an "as the title is touched" basis.

The scanning centers run by the Internet Archive have multiple scanning stations and require the constant through-put of material. Serial titles with the volumes and extensive runs provided the proper "food" to keep the scanning

"beast" in operation. In the natural sciences, with such a long history of publishing, there are many serial title runs that take up significant library-shelf real estate in most museum and botanical garden library collections. It was essential that the BHL partners identify serial runs that could quell the beast's appetite. The Serials Bidding list enabled a successful start to this project.

MONOGRAPH SELECTIONS

The MBLWHOI Library informatics team at the Marine Biological Laboratory and Woods Hole Oceanographic Institution developed an in-house monographic analysis tool. Even OCLC, which hosts arguably the most extensive metadata collection and has a research staff targeting library tool development, will say that machine de-duplication is not 100% successful. The BHL Monographic Deduper is not without its flaws, but it allows libraries to begin to select monographic titles to get them to the scanner.

The BHL Monographic Deduper is a Web application developed using the Ruby on Rails open source Web framework. Designed with the workflow of BHL librarians in mind, this tool makes use of the packlists created to accompany each shipment of items to the scanning centers. Each BHL library has its own workspace in the application to which packlists (in .xls format) can be uploaded. After discussion, the BHL librarians agreed to standardize the packlists to contain the following column headers: local ID, OCLC, title, author, volume, chronology, call number, publisher, and publisher place. While other columns may exist in the packlist, the standardized headers must exist for the tool's ingest process to work correctly. Upon upload, a packlist is parsed and all records are added to the database and a new entry is recorded in that library's work space containing the name of the packlist, the upload date and time, and options to view duplicates, show the entire packlist, or delete the packlist. Viewing duplicates for a given packlist will initiate five SQL queries against the entire database of previously uploaded packlists. Possible duplicate items are displayed in the following five buckets:

- Duplicates by OCLC and Volume
- Duplicates by OCLC only
- Duplicates by Title and Author
- Duplicates by Title and Chronology
- Duplicates by Title only

For each possible duplicate record, the option to view the suggested duplicate and its scanning library is given (see Figure 2). If a suggested duplicate is determined by a librarian to be a valid duplicate, that record can be deleted from the current packlist right from the Deduper's user interface. Once the

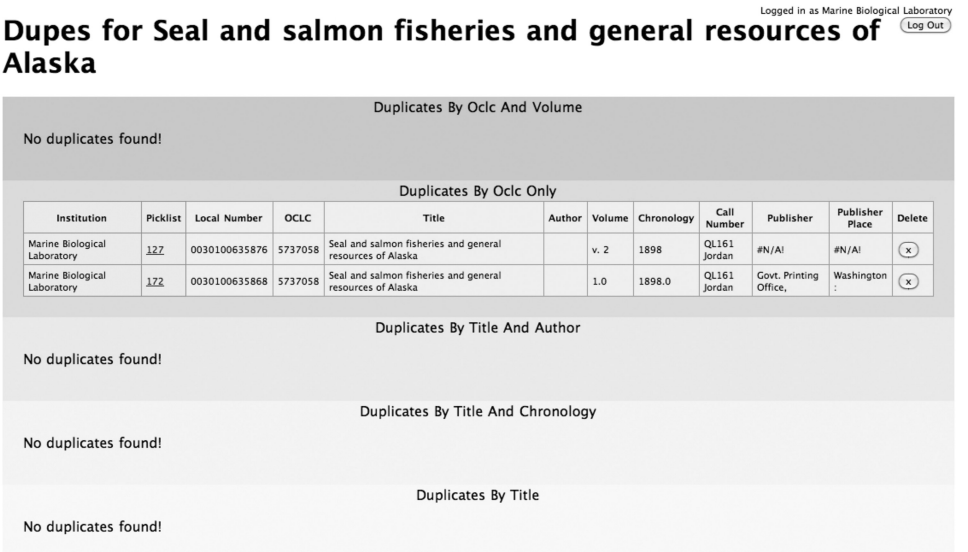

FIGURE 2 Results screen in Monographic Deduper.

list of suggested duplicates has been reviewed by a librarian and valid duplicates have been removed, an updated packlist can be downloaded in .csv format. As with most of the coding work done for BHL, the Monographic Deduper is made available as open source.[6]

But even with these tools in hand, it was the partnership and collaborative spirit that truly got all the BHL partners moving to get books to a scanning center. Common collegial communication became integrally linked with the use of the above serials and monograph tools, and subject selection took place: MBLWHOI Library would do sea creatures, NHM London would do general serial runs, AMNH would begin with birds, MCZ Ernst Mayr Library would begin with amphibians and reptiles, and the Smithsonian Institution libraries would begin with entomology. The botanical libraries worked out the areas of strengths and committed to scanning their host institution publications.

METADATA CHALLENGES

Serials are an interesting metadata challenge. It's well known that librarians either fall in love with the complexity of serials title changes, merges, splits, prediction patterns, publisher changes, societies; or they are annoyed by these changes given the work required to update catalogs. It became apparent that a number of BHL library staff actually have a love-hate relationship with serials: we love that they fill up a shipment to the scanning "beast" and have the complexity of coverage for the BHL until we hit one in a foreign

language that spans such a long time period as to have war interruptions, topic changes, and title changes that flip back and forth. What was surprising to most of the BHL staffers though, was how little is known about serials outside of the library world.

BHL staff met for a targeted meeting with Internet Archive programmers to explain the concept of volumes and issues, series and serials, and the idea of separate libraries binding the same title into different packets. Terminology used by various staff doing various aspects of a mass-scanning project was quite enlightening. Team members saw this development as a true sociological undertaking. Engineers and programmers think a "book" is something that has cardboard on either side—"cardboard to cardboard." Serials are bound into "books" and, therefore, the metadata describing the title of the serial is exactly the same for each of these books. This was seen as a warning sign by the catalogers participating in the project. With serial runs of hundreds of volumes, discovery of the proper volume post scanning is impossible with such a definition. Through these discussions we learned the lesson that any librarian who wants to try to explain volumes, issues, and numbers bound over a span of hundred years to someone who thinks in "cardboard to cardboard" should never try to convey the details over the phone. Whiteboards and face-to-face meetings with examples in hand are the only way to resolve these issues.

Once a clearer understanding of the metadata needs of traditional serials had been resolved, we tackled the following questions: (1) How do you transfer the metadata associated with a physical volume to the scanning center, (2) How is that information ingested into the Internet Archive digital system, and then (3) How should the data then be transferred, ingested, and displayed in the not yet fully developed BHL portal?

PACKING LISTS TO SCANNED BOOKS

WonderFetch (BHL would like to trade mark the phrase!) was developed very quickly by the Internet Archive once the concept was clearly understood. It accommodates the needs of serial holdings and as a bonus it has built in copyright clearance language. WonderFetch supplements the basic bibliographic description allowing each "book" to have the title description, specific volume, issue number, and date metadata. Though simple (not simplistic), it is extremely important for discovery and proper identification of serials.[7]

WonderFetch makes use of a spreadsheet application to compose URLs for each item in a shipment bound for the Internet Archive scanning center. Internet Archive engineers devised this as an efficient mechanism to populate their biblio database with both bibliographic and item-level data from library-generated packing lists. An Excel (or OpenOffice) spreadsheet packing

list is created by the library for each shipment of materials sent to the scanning center. The packing list contains data for each item in the shipment. This data typically includes library name, library catalog number (or some unique number to identify the correct bibliographic record within the library's integrated library system), barcode number, volume/issue/part designation, chronology, call number, title, author, date of shipment, and special notes or instructions to the scanner (see Figure 3). This data is copied to a WonderFetch spreadsheet template containing formulas and a selection of copyright statements. For each item in the shipment, the appropriate copyright statement (e.g., "Not in copyright," "Digitized with the permission of the rights holder," etc.) is selected, and the bibliographic, item-level data and copyright information are concatenated to create URLs. By clicking on the URLs formulated within the spreadsheet, bibliographic data is "fetched" from the library catalog via a negotiated Z39.50 connection, and item-level data (volume, number, date, etc.) are pulled from the spreadsheet. Thus, for serials items especially, essential metadata is extracted and stored in the meta.xml files for each volume, issue, or part. Items receive appropriate description in the BHL portal (see Figure 4).[8]

FRANKENBOOKS

Frankenbooks is the term used within the BHL refering to the digital version of a title that might have pages scanned and "stitched together" from different physical books. How to conceptually address the handling of this situation was debated within the group. The goal of getting the data locked on the printed page digitized, OCR'ed (i.e., optical character recognition applied), and made available to the wide user community seemed to support the idea of getting the title scanned. But the importance of the original source of the data is critical. As mentioned previously, in the taxonomic world, nomenclaturists value the date of the printed page. The exact date that a species is named is extremely important (see the history of T. Rex; Breithaupt, Southwell, & Matthews, 2005). It was decided that BHL will not allow for Frankenbooks because it was too important to be able to trace the digitized page back to the physical page with a specific date. If a title cannot be scanned because of missing pages, then that physical piece is rejected. Attempts are made to find another copy of the book at a partner library.

"Frankenserial" runs were deemed acceptable. Remembering the practicality of feeding the beast and the mission of providing access to our users, gaps in serial runs would have to be filled in by one of the partner libraries. Unlike a monographic title, serial runs are extremely difficult to find as complete sets in one library's holdings. Due to copyright restrictions and institutional policy, NH London will scan a title up to and including the year 1860, although if it can, through a due diligence process, obtain clearance

Institution	Item	Book ID	HOLLIS#	Barcode	Title Begin Date	Title End Date	Volume	Item Date	Call Num	Title	Author	Date Sent	Notes
MCZ	1		129615	32044106258254	1890	9999	v.12 (1900)	1900	ENT 2658	The Entomologist's record and journal of variation		23-Mar-2010	
MCZ	2		132137	32044110350568	1952	9999	v.46:no.3/4 (1997)	1997	NAC 4948	Nachrichtenblatt der Bayerischen Entomologen		23-Mar-2010	scan v.46:no.3/4 only
MCZ	3		5253936	32044110340809	1769	1829	1 Bd. (1769)	1769	QL406.M37 1769	Neues systematisches Conchylien-Cabinet	Martini, Friedrich Heinrich Wilhelm, 1729-1778	23-Mar-2010	
MCZ	4		5407791	32044114229735	1840	1841	text	1840-1841	Spec. Coll.	Die Versteinerungen des nord-deutschen Kreidegebirges	Roemer, Friedrich Adolph, 1809-1869	23-Mar-2010	
MCZ	5		5407791	32044114229743	1840	1841	atlas	1840-1841	Spec. Coll.	Die Versteinerungen des nord-deutschen Kreidegebirges	Roemer, Friedrich Adolph, 1809-1869	23-Mar-2010	

FIGURE 3 Typical packing list with examples of bibliographic and item-level data.

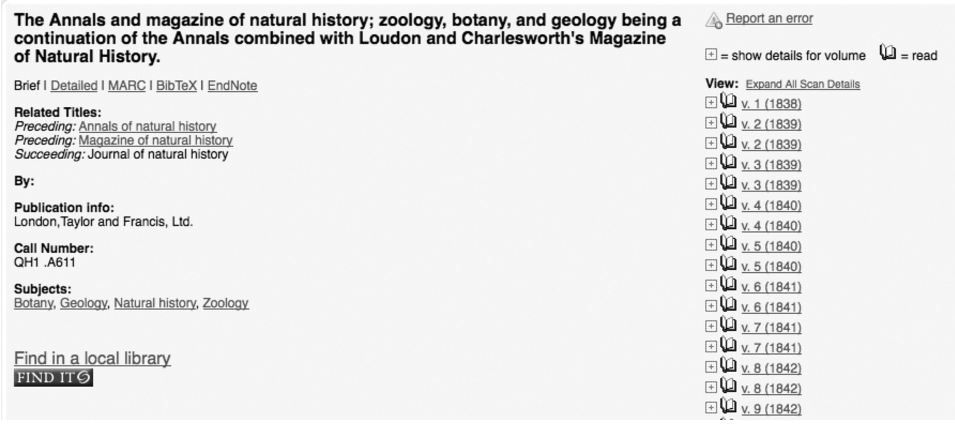

FIGURE 4 Serial record in BHL portal showing items with full enumeration and chronology data. Duplicate volumes indicate contributions from two libraries.

to scan a title, it will do so for titles beyond that date; MCZ Ernst Mayr Library will not scan beyond 1908 for non-U.S. publications. Therefore, the rest of the title is scanned by other BHL partners—copyright limitations for the partners allow scanning up to 1923. Missing volumes or volumes physically too fragile or otherwise not fit for scanning are requested from partners for filling in. These procedures have been extremely successful in enabling us to overcome the various random binding decisions of each of the BHL partners throughout their history.

FOLDOUTS

Older natural history serial titles include the "pop up window" of their day—the foldout. While the BHL scanning partner, the Internet Archive, had developed a successful book ("cardboard to cardboard") scanning workflow, they had a learning curve to tackle with respect to how to deal with older serials volumes with foldouts. Foldouts range from a simple one-inch folded extension of a page, to several feet in length and width built out of the binding of a volume. The Internet Archive scanning stations are designed to hold a book in a cradle and shoot each open page simultaneously using two overhead mounted cameras. Without a process in place to scan large scale foldouts, the BHL began scanning operations holding in reserve every volume that was published with foldouts. This required keeping track of titles or specific volumes that contained foldouts so that these volumes could be retrieved for scanning when the Internet Archive had the facility to do so. This situation contributed to numerous silos of metadata and, in some cases, separate shelving for waiting volumes. When Internet Archive developed

a scanning station for foldouts one year into the project, these reserve volumes were then scanned and previously skipped foldouts were digitally stitched into the originally scanned volumes. Thus, all of the data contained in the foldouts was accounted for. The monographic tracking system had to be updated to indicate that these previously "on hold" items were now complete. This is a notable example of moving ahead with a project before all of the essential tools and work flows have been tested and are in place. The pressure of moving the project forward was what was needed to resolve this issue.

QUALITY CONTROL

Quality control is significant for the undertaking of any digital initiative. A number of the main practical issues involved in any scanning project are quality control of metadata, images, scanned book volumes, and every nuance involved in the conversion of thousands of books into millions of scanned pages.

Questions arise, such as: What do you do if the book left your library to be scanned with one title attached to it on the spreadsheet, and the book ended up digitized attached to another title? What if the scanned page was blurry, or words were cut out of the image? These problems were addressed by the BHL librarians in various targeted ways. One method developed to examine scans was to take a statistical sampling of volumes scanned and compare the page by page quality of the scanned pages to the actual in hand paper copy. At the same time, captured metadata was examined as well. When metadata diacritics were a problem, this was communicated to other BHL librarians, and the Internet Archive staff was quickly informed of the issue. We have since solved the issue to ensure that discovery is not jeopardized due to metadata corruption. To achieve this coordinated effort, a number of quality control conferences were held, face to face, via conference calls, and via a dynamic wiki page interface.

TRACKING PROBLEMS

Initially, errors found by librarians and the public were reported and discussed via email. The volume of exponentially growing threads on many issues through this type of communication quickly became overwhelming. In a solid step forward, borrowing from the technology field, the BHL staff instituted a Web-based error tracking system for communicating metadata errors and quality control issues. The Gemini system[9] allows either a library user or a librarian to create a "ticket," which is then reviewed by a quality control librarian who assigns the ticket to the appropriate librarian at the

appropriate institution for issue resolution. The data associated with any issue and its resolution is preserved in an electronic issue resolution system for future reference. In practice, these types of tools have been in use by the technology field to track bug problems in programming but had not been adapted to the world of digital librarianship. This remarkable system is a technical-human interface tool that works well and will guide the resolution of many metadata issues for years to come.

POSTSCANNING METADATA WORK

BHL staff, like many catalogers, prefers that metadata be created once and repurposed as needed—"touched once" in essence. But, as documented above, there are many opportunities in the complex scanning work flow for errors, omissions, and unintended missing data to occur. In many cases it is typical that some postscanning editing needs to be performed by staff members. The BHL portal, through which the library's scanned content is viewed, is our only access point to book-associated metadata. As the BHL was developed using open access technology solutions, it is not as robust as most high-end integrated library systems. It lacks the traditional search and editing functionalities a librarian would like to see. The current BHL portal technical infrastructure was established as a viewing system and inventory of scans enabling computer-to-computer uses, such as data mining, and connection to services such as the Taxon Name Finder (also discussed in detail below). Taxon Name Finder, an application developed at the MBLWHOI Library, mines OCR'ed text to return all the scientific names appearing on each page of literature in the BHL. This lack of metadata editing and field controls has been challenging to learn how to resolve.

Described above, a Frankenserial is the result of two or more BHL partners contributing scanned volumes of a single journal title. After scanning, serial records need to be merged into one metadata record and the volumes must be sequenced. For example, SIL contributes volumes 1 through 66 of *Tijdschrift voor Entomologie*. Volumes 67 through 142 are picked up by MCZ Ernst Mayr Library and so forth. With each institutional contribution to this title, a MARC record is added to the portal. In this example, two records for the same serial title now reside in the database. There should be only one BHL record with all contributed volumes so that users do not encounter multiple hits when searching for a title. Administrative editing allows for the merging of the titles and sequencing of volumes. Quick roll out of new tools and administrative editing capabilities has allowed the BHL to accommodate the formerly unconsidered Frankenserials.

Other tweaks and adjustments made quickly to the editing functionalities included the need to provide proper citation resolving and discovery points. Occasionally, the host library does not have all the information regarding

the enumeration and chronology of each volume scanned. In such cases the WonderFetch spreadsheet had incomplete information. Each one of those volumes that are added to the BHL collection needs attention to indicate the proper volumes and years associated with the scans. Other typical edits include adding access points for additional authors that are requested by the users of BHL but were lacking in the original bibliographic description, page number (described below), and other types of mergers of metadata records to point to complete sets of titles.

THE MONOGRAPHIC SERIES PROBLEM

Monographic series present a unique problem. For example, the Field Museum contributed their publication *Fieldiana* to the BHL. Librarians at the Field Museum have analyzed each number (issue) and have thus contributed to the BHL, a fully analyzed monographic record for each number of the series. However, the database should also have a serial record for *Fieldiana* so that users may search that title and browse all numbers of the series. Toward this end, BHL programmers added portal editing functionality that allows librarians to associate a scanned item (e.g., an issue, number, or volume) with more than one bibliographic record. Using the example of *Fieldiana*, a serial record can be uploaded to the portal, after which all scanned numbers may be associated and displayed with that serial record while remaining linked also to their original monographic records. Likewise, volumes originally loaded under a serial title such as *Memoirs of the Museum of Comparative Zoology* may be subsequently associated with monographic records. Associating scans with two or more titles is accomplished manually and is part of the work flow for monographic series. Thus, volumes of a monographic series may be discovered by a search for the serial title or for the individual monographic titles.

PAGINATOR

As a title is scanned by the Internet Archive scanning technician, page numbering is asserted. When doing the initial quality review of the scans, the technician can indicate the beginning page number and have the system auto generate the pages numbers. With some attention to detail, most of the traditional printed materials are "published" live to BHL with page numbers indicated. But the BHL project has many titles that do not fit into the traditional mold. Users of the site will notice that the system attempts to determine whether the page being viewed is text or images, although a page number is not given. Assigning page numbers within a text is a labor-intensive human intervention solution. Missouri Botanical Garden program staff have

developed a client application called the Paginator. There is a Web version as well for those with administrative editing privileges on the BHL portal.

After logging into the administrative portion of the BHL, a scan is chosen and opened. Each page is then opened and a number is assigned. In the end of the process, each page has been assigned a literal page number in the sequence of scans and the page assertion for resolving citations. For example, the sixth page in the scan might actually be page iv of the book since initial scans will be the cover, title page, verso of the tile page, et cetera. An important aspect of pagination is correctly asserting page types—another function of the Paginator. For example, pages may be designated as text, illustration, issue start, foldout, index, et cetera, assisting navigation and discovery within a digitized work. This is especially important for taxonomic literature because investigators are often seeking specific illustrations of organisms. The Paginator allows free-text description of pages in addition to asserting page types. Thus, illustrations or foldouts may be described in some detail if so desired. Again, this is a manual process and thus very labor-intensive.

OPEN URL

BHL has focused development efforts to provide access to and resolving of citations. The BHL's OpenURL query interface is available at http://www.biodiversitylibrary.org/openurl. Both OpenURL 0.1 and OpenURL 1.0 queries are supported. The table summarizing the parameters that are accepted by the OpenURL 0.1 and 1.0 query interfaces is available at http://www.biodiversitylibrary.org/openurlhelp.aspx. By default, the query interface will (if possible) redirect to the BHL page containing the citation described by the query. If more than one possible citation is found, the query interface redirects to a page from which the appropriate citation can be selected. There are several additional ways that results from the query interface can be returned: JSON, XML, and HTML. If results are returned as JSON, a callback function may also be specified by adding a "callback" argument to the query.

TAXONOMIC INTELLIGENCE

BHL uses TaxonFinder,[10] a taxonomic intelligence tool developed by collaborators at uBio.org,[11] to locate and identify scientific species names within the text of digitized books. This names-based index is an incredibly valuable tool for research on organisms and specific genus and species. It is easily incorporated into external Web sites. A full bibliography of all the titles contained in BHL that have a specific species named can be generated by a stable url that launches the species name search. To easily link

into a list of all pages containing a given scientific name, follow the pattern http://www.biodiversitylibrary.org/name/*Scientific_name*. The example below will provide an up to date result that includes the most recent scans of material that mention the typical goldfish.

> Example: Typical goldfish scientific name *Carassius auratus*
>
> http://www.biodiversitylibrary.org/name/Carassius auratus

For computer to computer calling the name service has been established on the BHL portal. The name services are XML-based Web services that can be invoked via SOAP or HTTP GET/POST requests. Responses can be received in one of three formats: XML wrapped in a SOAP envelope, XML, or JSON.[12]

BHL has yet to resolve the need to translate the citation abbreviations that taxonomists use. With the help of TaxonFinder, BHL exposes to the users a way to access the specific species. The particularly challenging issue of resolving journal title abbreviations has yet to be solved.

TOWARD A GLOBAL BHL

This paper provides an overview of the BHL project, the support it offers to vital work in the taxonomic field, and library metadata challenges and advancements. The paper explains metadata work flow that parallels the work flow of the physical items from shelf to scanner and beyond to delivery of the information to the Web. Tools, including the Monographic Deduper and the Serial Bid list, were created to deal with the issues surrounding choosing materials to be scanned. Work flow to enhance the data supplied to the BHL portal included the development of WonderFetch and the Paginator. The work conducted so far has been successful because of the collaborative spirit among the team members. We have accomplished the creation of a body of digital material already used in the daily work of scientists. The availability of OCR text and computer-to-computer interface with tools like TaxonFinder has made the data within the literature usable. We have identified areas to approach next to ensure quality of the collection, ease of discovery, and reuse of the data locked on the pages of the centuries-old literature and to provide the metadata to a wider community.

The BHL project has been expanding globally because biodiversity is an increasingly important topic. There are many converging elements (environmental, climatic, biotic, and agricultural) that are turning the often overlooked and underfunded science of systematic taxonomy into a globally relevant topic of interest.The quick successes of the BHL in providing access to vast amounts of taxonomic literature have stimulated others around the world to contribute to the project. First out of the gate was BHL-Europe in

May 2009. This European Union-funded project consists of 26 institutions from across the EU. Partnering with BHL "classic," BHL-Europe is providing data storage and additional technical development.[13]

In late 2009, the Chinese Academy of Sciences and BHL signed an agreement for BHL-China. BHL-China describes itself thus:

> Chinese Biodiversity Heritage Library (BHL-China), the pre-research project funded by the Biodiversity Committee, Chinese Academy of Sciences, is aiming to, through collaboration with BHL (Biodiversity Heritage Library) and in conjunction with other institutes (colleges) on biological research, jointly build a network platform for BHL-China; through the comprehensive collection, scanning, extraction of the essential biodiversity related literature and the systematical arrangement of the important biodiversity (early focus on botany) literature, to establish an easily-searchable and communitive network platform, while to make the data API compliant and therefore provide documentation data services to biodiversity (including EOL China nodes, Chinese Virtual Herbarium, etc.) and other related research fields.[14]

Late 2009 and early 2010 saw the beginnings of even more globalization of BHL. An initial meeting was held at the Bibliotheca Alexandrina for an Arabic-language BHL. In early 2010 organizational meetings were convened for a BHL node based in Brazil. Discussions are underway with the Atlas of Living Australia for BHL-Australia.

The work pursued via BHL is forging new groups and is modernizing the solution discussed at the British Museum in 1847, when Charles Darwin and a group of scientific luminaries commented: "The cultivation of natural science cannot be efficiently carried on without reference to an extensive library" (Darwin et al., 1847). Today, Darwin's "extensive library" is an increasingly global virtual library designed by the most forward thinking librarians, scientists, and informaticians. This library is freely available to researchers around the world, and at the service of those studying life in its myriad forms.

NOTES

1. International Commission on Zoological Nomenclature's International Code, http://www.iczn.org/iczn/index.jsp

2. Article 8 and recommendations in Article 8 of the International Commission on Zoological Nomenclature's International Code, http://www.iczn.org/iczn/index.jsp

3. Some fields, such as botany, tend to reference an earlier edition dated 1753.

4. Encyclopedia of Life, http://www.eol.org

5. CakePHP, http://cakephp.org/

6. The open source code for the BHL Monographic Deduper can be found at http://github.com/woodshole/BHL-dedup

7. Details of WonderFetch and other workflow issues are documented here: http://biodivlib.wikispaces.com/Workflow

8. A slight variation on the theme of WonderFetch at each BHL-participating library has been adapted to ensure proper workflow. SQL databases and MARC data dumps have been implemented.
9. Gemini is a product of CounterSoft (http://www.countersoft.com/home.aspx)
10. TaxonFinder is available at http://www.ubio.org/index.php?pagename = xml_services
11. uBio.org services are available at http://www.ubio.org/
12. A full description of these services is available at https://docs.google.com/Doc?id=dgvjvvkz_1×5qbm3
13. BHL-Europe, http://www.bhl-europe.eu
14. BHL-China, http://www.bhl-china.org/cms/node/25

REFERENCES

Breithaupt, B. H., Southwell, E. H., & Matthews, N. A. (2005). In celebration of 100 years of tyrannosaurus rex; manospondylus gigas, ornithomimus grandis, and dynamosaurus imperiosus, the earliest discoveries of tyrannosaurus rex in the west; geological society of america, 2005 annual meeting. *Abstracts with Programs—Geological Society of America*, *37*(7), 406.

Darwin, C. R., Murchison, R. J., Buckland, M., Egerton, P. G., Greenough, G. B., & Owen, R. (1847). Copy of memorial to the first Lord of the Treasury [J. Russell], respecting the management of the British museum. *Parliamentary Papers, Accounts and Papers*, *24.253* (Paper No. 268), 1–3.

Gray, J. E. (1850). *Catalogue of the specimens of mammalia in the collection of the british museum*. London, UK: Trustees of the British Museum of Natural History.

Moritz, T. (2005). *Library & laboratory: Civil union??? (slide 26)*. Unpublished manuscript. Retrieved from http://barcoding.si.edu/LibraryAndLaboratory/3–11_Moritz.pdf

Sherborn, C. D. (1932). *Index animalium*. Cambridge, UK: The Trustees of the British Museum. Retrieved from http://www.archive.org/details/sil34_02_29