Biodiversity Informatics, 12, pp. 58-75

CROWDSOURCING NATURAL HISTORY ARCHIVES: TOOLS FOR EXTRACTING TRANSCRIPTIONS AND DATA

KATHERINE MIKA*, JOSEPH DEVEER, AND CONSTANCE RINALDO
*Ernst Mayr Library, Museum of Comparative Zoology, Harvard University, 26 Oxford St, Cambridge, MA 02142 USA. *Corresponding author: kmika@fas.harvard.edu*

Abstract.—This paper surveys the landscape of current, successful, and innovative crowdsourcing platforms for obtaining full text transcriptions and structured datasets hidden in manuscript items in the Biodiversity Heritage Library. Transcribing manuscripts are optimal tasks for crowdsourcing programs because they require intellectual engagement and thoughtful decision making to produce meaningful content. By offering full text transcriptions, digital collections are opened up to new types of searching, sorting, categorizing, and pattern finding. Research derived from these new datasets can illustrate changes over time across much larger magnitudes of collections and types of information resources. A targeted analysis of methods, tools, and programs for crowdsourcing manuscript transcriptions describes the challenges and opportunities in developing a project that produces machine readable facsimiles and can support structured data extraction from natural history libraries and special collections content.

Key words.—field notes, transcription, data capture, primary biodiversity data, crowd-sourcing

The Biodiversity Heritage Library (BHL) is a global collaborative digital library established in 2006, with a mission to improve research methodology by making biodiversity literature openly available and inspiring discovery through free access to biodiversity knowledge (Gwinn and Rinaldo 2009). The library is adding manuscript collections to its collection in order to expose hidden and often encrypted observation and occurrence data. These data are collected from machine readable text transcriptions and document the environment, climate, and biodiversity over wide temporal and geographic domains. BHL in partnership with the Library of Congress' National Digital Stewardship Residency program has made it a priority to add text from manuscript transcriptions to enhance the quality and completeness of machine readable biodiversity data in the digital library portal and developer interface.

Manuscript items contain a wealth of occurrence data which cannot be investigated without reliable transcriptions. A general survey of transcription utilities revealed that there are many types of projects with dramatically different goals and user needs. Therefore, a targeted survey will provide better guidance on methods, tools, and programs that can support structured natural history data extraction as well as free text transcription. Current crowdsourced transcription platforms build on the successes of earlier projects and learn from their mistakes and challenges. Because a consortium digital library depends on contributions from its members, a transcription platform must be easy to implement, generate output that is compatible with various types of transcription files, and function within several layers and types of encoding. The selected program will likely be used by libraries to transcribe local items in addition to content for BHL. Several successful platforms have been identified as potential tools with which to develop a cohesive transcription program for BHL.

The BHL digital library portal is built on a relational database that defines and indexes relationships between Titles, Items, Pages, and Segments and their administrative, structural, and descriptive metadata (Figure 1). As scientists increasingly turn to computational methods to answer large scale questions about

the natural world, BHL plans to update its data model to provide better access to collection data and taxonomic, bibliographic, and descriptive metadata. Natural history libraries are mandated to preserve and maintain the published literature that is critical to the discovery, revision, and naming of life. Over time, nomenclature specialists generated species citations according to discipline and sub-discipline specific abbreviations independent of related species projects or the expertise of those collecting and providing accessto taxonomic literature. Concurrently, librarians developed and implemented metadata policies without consulting taxonomists or the scientists using the information (Pilsk et al. 2010). This has resulted in a corpus of heritage literature and special collections that preserves taxonomic information and citations but is relatively inaccessible to the modern scientist. In order to mitigate this, BHL relies on the Global Names Recognition and Discovery Service[1] to mine and index the literature's text for stringsof potential Linnaean binomials, resolve them to existing taxonomic databases, and link the strings to pages in the BHL portal.

FIELD NOTES

The Biodiversity Heritage Library is adding unpublished manuscript materials from its contributors in the form of collector's field notes, diaries, and correspondence. Because Optical Character Recognition (OCR) in its current state of development does not accurately render handwritten materials, manual transcription of digitized manuscripts is essential for the creationof text files for each page (Kalfatovic and Rinaldo 2016). Natural History field notebooks and diaries are iconic symbols of scientific work and are important as a sequential record along with correspondence. Notebooks and correspondence showcase individual work but also provide insight into the personality of the writer and events of the time. Natural history, defined as biodiversity and environment of a region, is part of the cultural heritage that characterizes a nation or region and informs policy making (Europe 2005; Rinaldo and Smith 2014). The recorded information in field notes and correspondence also reflects changes in the philosophy of the study of natural history through time.

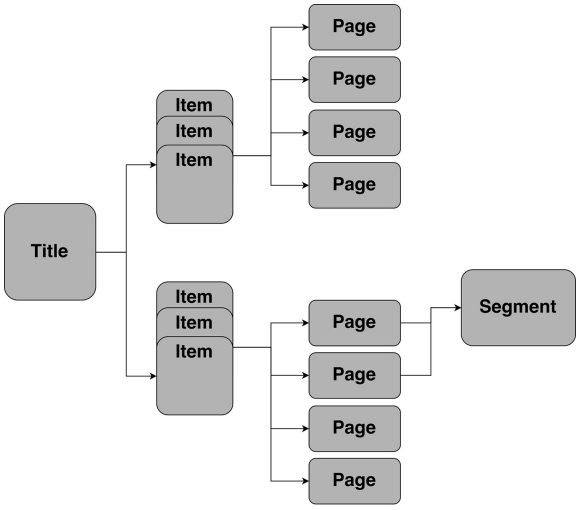

Figure 1: Adapted from BHL Data Model to illustrate how Titles, Items, Pages and Segments are related in the BHL portal[2].

Field notes and correspondence contain a wealth of raw scientific data including unpublished observations, species occurrence records, habitat descriptions, climatological data, phenological recordings, sketches, weather reports, and travel narratives: these records are primary source data at its most raw and unevaluated. Historical collections of field notes may be the only documentation of a scientist's thought processes, ideas, and observations, particularly if only some of the material was ever published (DeVeer, Rinaldo, and Ford 2013). When researched over time, biologists' field notes often document local environmental conditions and may help to identify gradual changes in number and presence of species over

[1] http://gnrd.globalnames.org/.

[2] https://github.com/gbhl/bhl-us/blob/master/Documentation/DataModel.

many years. Comparisons of species lists and descriptions of environmental conditions with current conditions provide valuable information about landscape and environmental changes and may pinpoint historical changes such as for species migrations. Digitized field notes provide an opportunity to review historical, cultural, and weather events of the times described in these documents in comparison with current conditions.

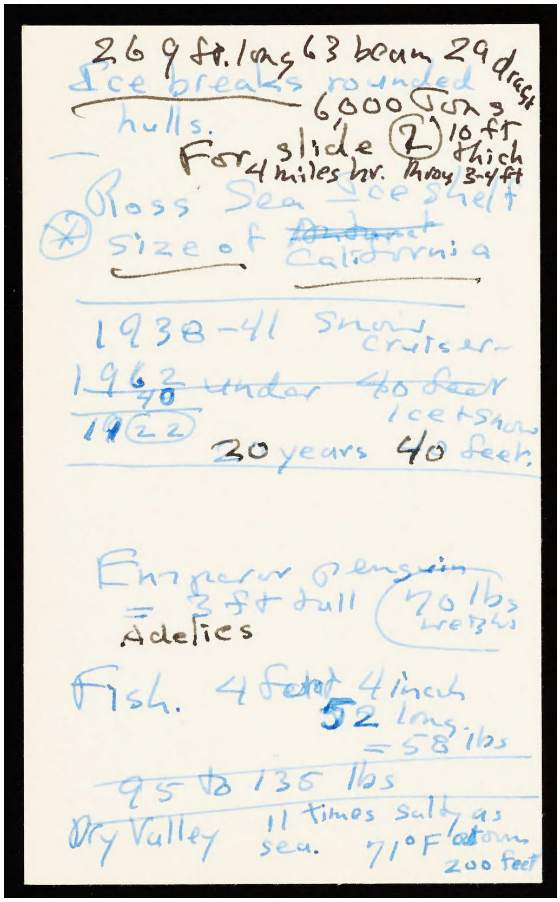

Figure 2: Excerpt from the Waldo L. Schmitt Papers that contains valuable observation and meteorological data and illustrates the creator's dynamic handwriting (Schmitt 1962).

Extracting specific information from handwritten documents using OCR is difficult, if not impossible. Some notebooks are crowded with writing in many different directions in an attempt to use as muchof the blank page as possible (Figure 2). To index this information and make it discoverable, machine-readable text files for manuscripts are needed in the BHL database so that text files can be produced for indexing as they are for published items such as books and journal volumes. The ultimate goal for BHL is to mine the vast amount of biodiversity data locked away in field notes, correspondence, images, and published texts and make the data accessible from the BHL corpus. Available information in BHL for a given species would then potentially include the original description of that species in the published literature with a link to an account of the field collection of the type specimen used as a basis for the description and to images in texts. Eventually, links to museum specimens and other non-literature-based data and objects will collocate all the elements of biodiversity information.

CROWDSOURCING

Digitized manuscript items are often hidden and inaccessible in digital libraries because their descriptive metadata is frequently minimal and their unique content is not discoverable without a machine-readable facsimile. Indexing transcriptions facilitates discovery of historical records and improves catalog search results. By offering full text transcriptions, digital collections are opened up to new types of searching, sorting, categorizing, and pattern finding. Research derived from these new datasets can illustrate changes over time across much larger magnitudes of collections and types of information resources. This is particularly important when considering biodiversity heritage literature and archives collections due to their significant value in documenting species occurrences, botanical observations, climate patterns, and meteorological events. Transcriptions facilitate the manipulation of this data and support research that extracts knowledge from formal and informal collecting and observation events. Transcription projects for collections are time consuming, intellectually intensive, and expensive for an organization to

facilitate. Crowdsourcing has been identified as a sustainable model for generating transcriptions for large collections and institutions with diverse holdings, and may improve data collection from a diverse range of users to enhance descriptive metadata.

After reviewing various types of collaborations and crowdsourcing in libraries, Sally Ellis argues that librarianship is strengthened by "cooperation with, and contributions from, users; relationships among libraries, archives, museums (LAM) and other information institutions; and the use of emerging technologies to facilitate these associations and interactions" (Ellis 2014). Crowdsourcing ensures a future for libraries, museums, and archives because it solves problems, strengthens collections and communities, and engages users. GLAM institutions (galleries, libraries, archives, and museums) are well prepared and appropriately situated to implement successful crowdsourcing projects due to their commitment to and history of facilitating public engagement. While discovery systems, online catalogs, and Web 2.0 tactics have been widely adopted to enhance remote access to collections, cultural heritage institutions are often reluctant to cede control of content selection, description, discovery, and use to digital volunteers (Holley 2010). Collaboration across institutions and users enhances access, facilitates use, and strengthens collections. Daren Brabham defines the term as an "online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals" (Brabham 2013). Tech companies, businesses, and academic research institutions have embraced crowdsourcing as a legitimate tool for generating new products, content, and operations, as well as for public relations and marketing purposes (Brabham 2013).

The Internet's speed, reach, temporal flexibility, anonymity, interactivity, and convergence brings people into conversation with each other, lowers barriers to information by creating easier access to professional bodies of knowledge, increases access to useful tools, and enables an online participatory culture (Brabham 2013). By externalizing transcriptions of manuscript items, we can leverage the collective intelligence and wisdom of crowds and exploit a large and diverse set of skills, tools, and ideas to bear on archival materials and special collections. The Internet encourages ongoing co-creation of new ideas in which content is generated through a mix of bottom-up (from the people) and top-down (policy-makers, businesses, and media organizations) processes (Brabham 2013). GLAMs are ideal institutions to encourage and utilize crowdsourcing initiatives due to their unique placement at the intersection of these processes. Libraries and cultural heritage institutions have the advantages of mission statements and codified ideologies dedicated to enriching knowledge as well as the organizational structures to mobilize, energize, and capitalize reciprocally on the capabilities of its users. This symbiotic relationship is not only mutually beneficial, but is likely one of the spaces in which GLAMs can thrive in the digital age.

Biodiversity research has a strong background in relying on non-scientist community members to collect data. These Citizen Scientist programs and the resulting data are understood as a "public good that is generated through increasingly collaborative tools and resources while supporting public participation in science and Earth stewardship" (Dickinson et al. 2012). Tracking and understanding biodiversity at varying scales requires fine-grain data to be collected over regions and continents, years and decades. Professional scientists alone are not generally capable of delivering the volume of data, analysis, and interpretation needed to support large-scale biodiversity research questions (Theobald et al. 2015). The study of sweeping patterns in nature requires vast amounts of data to be collected across an array of locations and habitats over span of years and often decades (Bonney et al. 2009) (Figure 3).

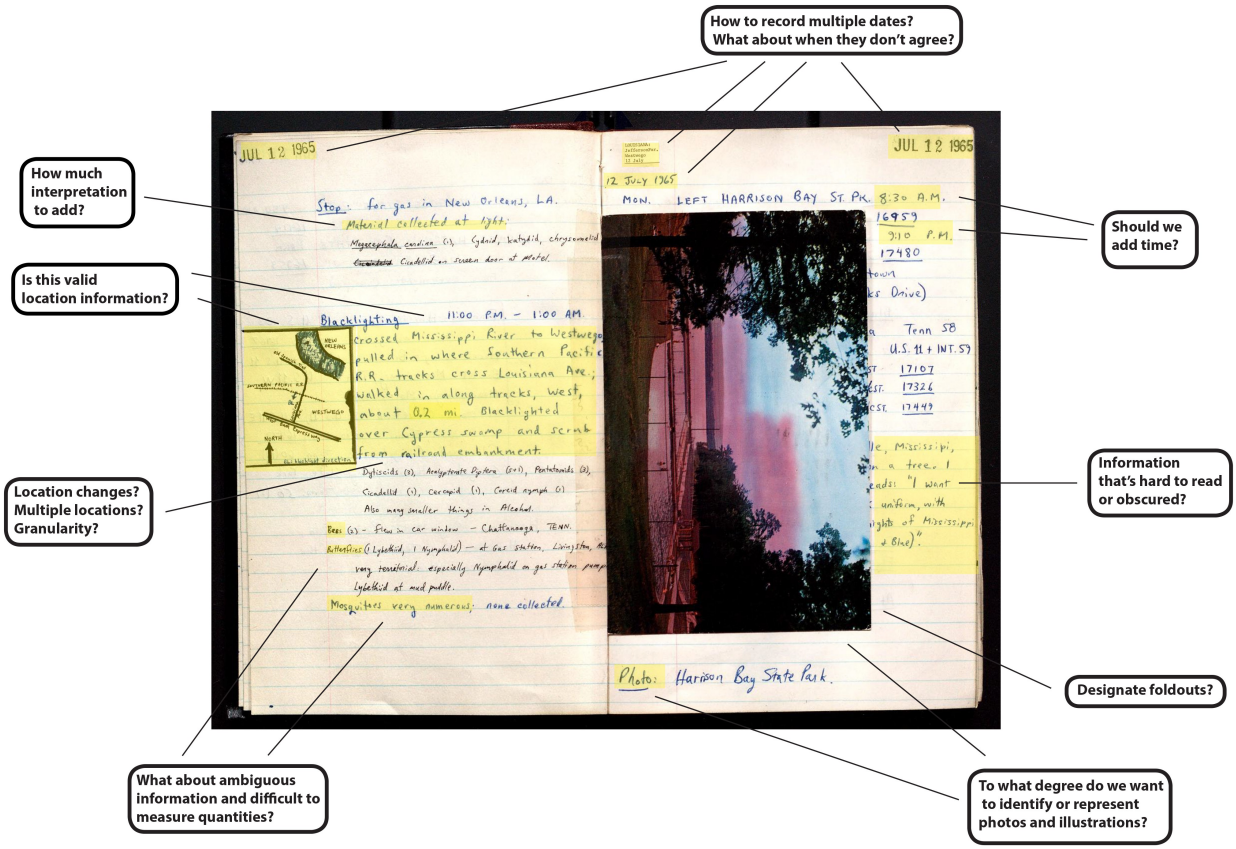

Figure 3: Excerpt from the *Robert E. Silberglied Papers, 1960-1982* illustrates the complexity of data capture in field notes and manuscript content. This page includes varying date and location information, Latin and common scientific names, occurrence and collection data, a map, and a photo foldout among other potential data points (Silberglied 1965).

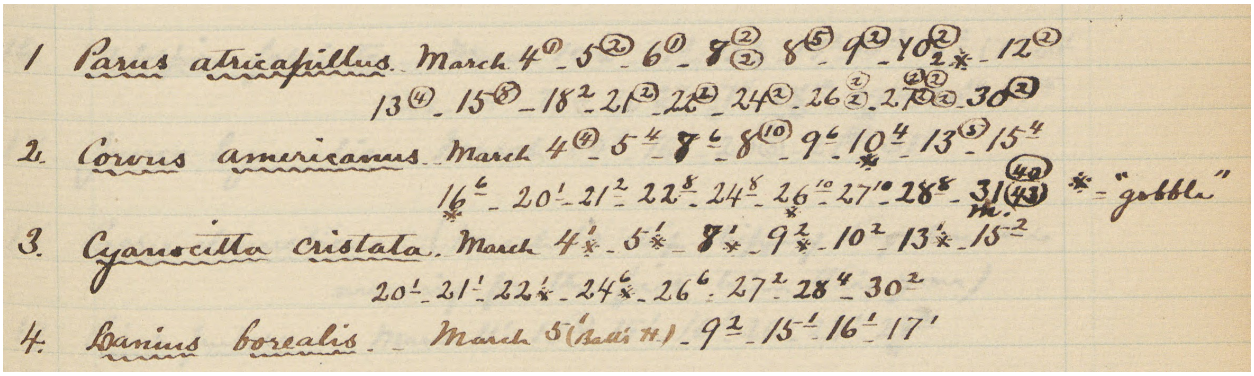

Figure 4: List of observed species from ornithologist William Brewster's field notes that illustrate the transcription difficulties in deciphering shorthand and abbreviations that are especially prevalent in scientist's notes (Brewster 1892).

Transcribing field notebooks and manuscripts, correcting OCR output, and curating collections occurrence data are optimal tasks for crowdsourcing and citizen science programs because they require intellectual engagement and thoughtful decision making to produce meaningful content. Collecting every potential data point from all possible documents generates datasets that are difficult to understand and obscured with irrelevant information. Crowdsourcing applies standardized processes for transforming content germane to scientific inquiry and collections' scopes into machine readable data points formatted for computational research processes.

Crowdsourcing transcriptions can be understood as a method of gathering data over wider geographical and temporal spaces. Field notes become powerful and rich sources of biodiversity information when the existing data is transformed into machine readable format. By transcribing and generating structured datasets from field notes, scientists of yore can be recruited for current research projects. BHL's content spans hundreds of years and the entire globe, creating a potentially vast pool of observation data that can inform current research (Figure 3). In the same way that science departments have turned to public participation to enlist the community in creating scientific knowledge, crowdsourcing transcriptions creates global networks that can generate data to be analyzed for population trends, range changes, shifts in phenologies, and more (Dickinson and Bonney 2012).

The crowdsourcing platforms discussed in this paper are the tools the Biodiversity Heritage Library is considering in order to develop a consortium-wide standard for extracting data from digitized items to improve the discoverability of hidden collections. The NDSR BHL transcription work addresses similar goals to the Art of Life (Rose-Sandler 2012) and Purposeful Gaming (Rose-Sandler 2015) projects that sought to enrich the metadata of items to better facilitate access to collections. The work of Art of Life sought to "liberate natural history illustrations from the digitized books and journals in the online Biodiversity Heritage Library through the development of software tools for automated identification and description of visual resources" (Rose-Sandler 2012). Images in BHL include page level structural metadata, facilitating navigation by human users and citation resolvers, but they lack sufficient descriptive metadata to enable dynamic filtering and inquiry. The Art of Life project built new software tools and algorithms to automatically identify illustrations found within the text pages of the BHL corpus and push those illustrations to crowdsourcing environments like Flickr and Wikimedia Commons for their description.

Similarly, full text searching of texts is significantly hampered by poor output from OCR software, and historic literature has proven to be particularly problematic because of its tendency to have varying fonts, typesetting, and layouts that make it difficult to accurately render. Purposeful Gaming was developed to identify a method for quick and efficient harnessing of large numbers of users to review and correct particularly problematic works by presenting the task as a game. Both projects improve the discoverability of and access to digital texts by enriching descriptive metadata for items at the page level to support full-text searching, data mining, and markup of content in BHL collections. The NDSR transcription project complements Art of Life and Purposeful Gaming by developing a similar method for generating machine readable content that will enhance access to handwritten text, a final category of "hidden content" in BHL.

TRANSCRIPTION

In the realm of manuscript transcription crowdsourcing platforms there are essentially two types of projects: record-based and document-based (Brumfield 2012). Record-based projects are more closely aligned with Citizen Science crowdsourcing because they seek to extract tabular data from handwritten materials (Figure 4). The output from these

projects can easily be stored in databases and searched, sorted, and categorized for findability and disseminated via APIs (Application Programming Interfaces). Content in these items is usually transcribed in online forms that structure the data in appropriate schemas depending on the project goals and the types of relevant information. Users and creators of these projects generally understand in advance what type of data is going to be produced from their collections and the research and analysis the data needs to support. Record-based projects are also not usually interested in representing the content of an entire document.(Brumfield 2012) Structured datasets generated from projects like Old Weather (Blaser 2014), FamilySearch Indexing (Holley 2010; Hanse et al. 2012), and the North American Bird Phenology Program and other specimen transcription citizen science projects (Miller-Rushing,Primack,andBonney 2012; Dickinson and Bonney 2012) follow this record-based strategy in order to support research over very broad geographic and temporal scales.

Document-based projects, in contrast, produce transcriptions that try to replicate the original digitized item—all aspects of it—as closely as possible, and in a machine-readable format (Brumfield 2012). Managers of these types of projects are usually interested in a full text output (encoded or unencoded) that records all of the content on a page. Platforms generally invite volunteers to type their transcriptions into a free-form text box that can support some markup conventions and may include toolbars to encourage consistency (Brumfield 2012). There are some encoding schemas and conventions that make it easier to communicate non-textual information digitally, but the kinds of markup for these types of projects are idiosyncratic at best, and do not adhere to a semantic standard or professional best practices. Many digital humanities and archives institutions are rapidly adopting TEI (Text Encoding Initiative) metadata schema for transcription markup, and while it is a good option that structures datafor display effectively (line breaks, annotations, insertions, deletions, etc.), it is fairly limited in its ability to digitally represent informational content outside traditional archives scopes, like natural history.

The Biodiversity Heritage Library is considering combining record-based and document-based transcription in order to represent all of the relevant informational content in a document in addition to extracting specific structured datasets that are of particular use to its researchers and users (Figure 5). BHL users include staff members at consortium institutions, other digital systems that harvest BHL content and data, and individuals that access items at all levels. The Encyclopedia of Life[3], the Global Biodiversity Information Facility[4], BioStor[5], and the Global Names Architecture[6] are systems users that link to BHL's content, often via taxon names, for bibliographies and foundational literature (McClanahan 2017). Georeferencing and including transcribed field notes and archival collections will further enhance interoperability and connections between types of biodiversity data. In addition to traditional access, individual users pull datasets from BHLs publicly available APIs in order to mine contents (primarily) for species occurrences.

Currently, the BHL data model (Figure 6) describes titles, pages, and segments by defining relationships with keywords, authors, and scientific names. Authors and Keywords can describe Segments or Titles, but not Pages or Items. Scientific Names are attached to pages, but not Titles, Items, or Segments. For Field Notes and other manuscript collections additional metadata tables such as Locations, Dates, added Personal Names and Corporate Bodies, and perhaps Expeditions could be added. Currently Dates are defined at the Title level via an imported MARC record, and Expeditions are

[3]http://eol.org/.
[4]https://www.gbif.org/.
[5]http://biostor.org/.
[6]http://globalnames.org/.

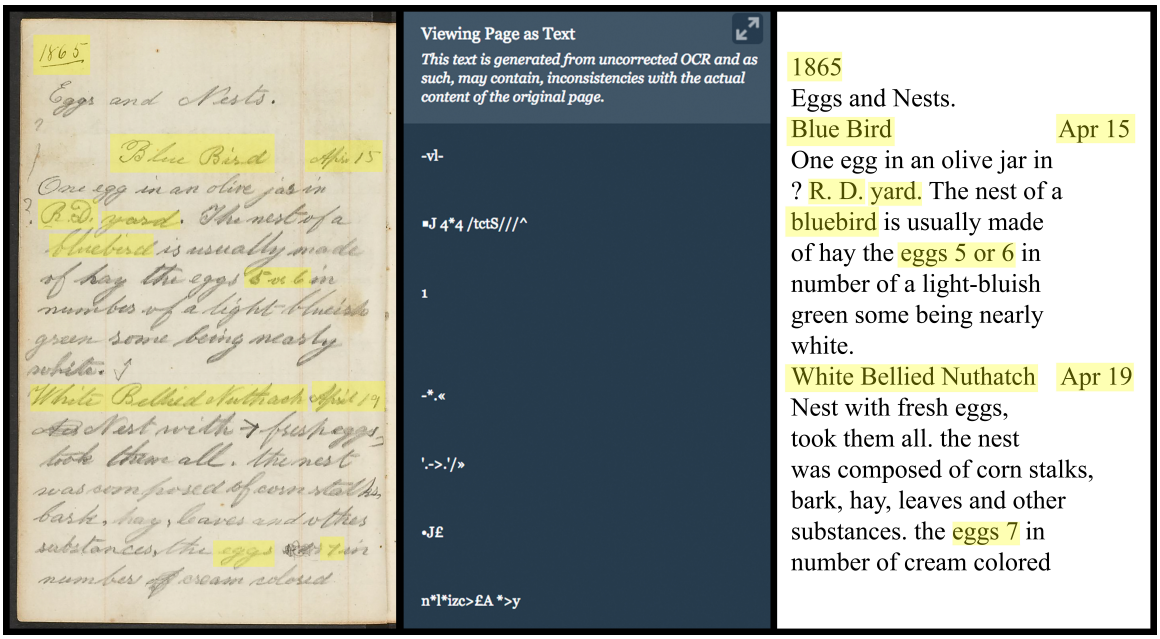

Figure 5. Image features manuscript page from ornithologist William Brewster's Diary (1865), the automatically generated OCR, and its transcription with potential species occurrence and description data points that BHL would like to capture highlighted (Brewster 1892).

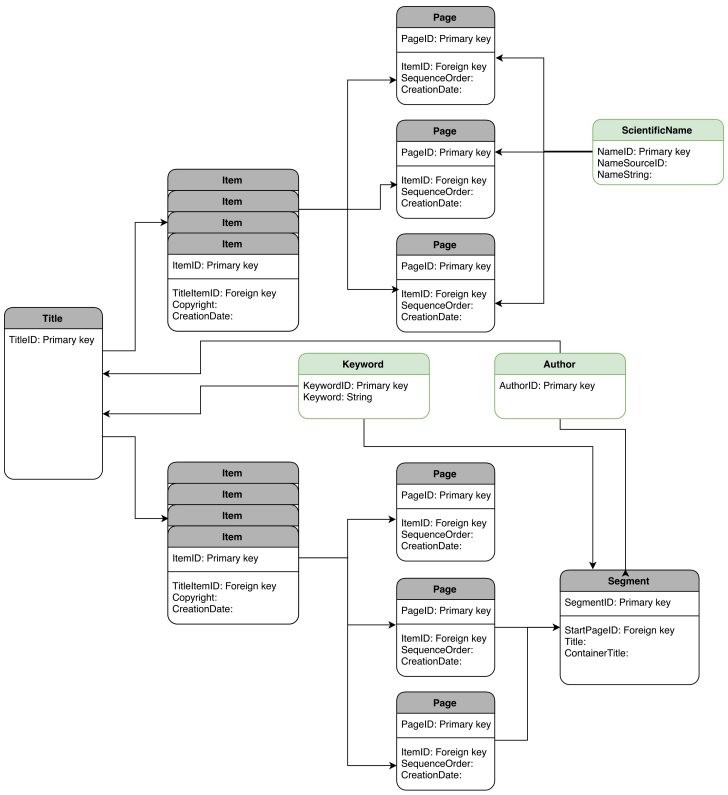

Figure 6. Illustrates the relationships between Titles, Items, Pages, and Segments, and associated Keywords, Authors, and Scientific Names. https://github.com/gbhl/bhl-us/blob/master/Documentation/DataModel.

linked through curated bibliographies. BHL will need to consider where to add new tables, what description level they should be at, and whether to add or modify description levels of the existing metadata tables.

EARLY TOOLS

Two early platforms, Scripto[7], developed by the Roy Rosenzweig Center for History and New Media at George Mason University (RRCHNM) and Transcribe Bentham[8] shaped the landscape for successful transcription projects. Archivists at RRCHNM synthesized Web 2.0 tactics and crowdsourcing successes from outside of the cultural heritage and GLAM arena to build an application to transcribe the Papers of the War Department (PWD)[9].

In addition to leveraging Max Evans' concept of commons-based peer-production (Evans 2007) the PWD project hoped to use transcriptions to enhance the collection's findability. The RRCHNM tool allows users to "easily submit those transcriptions and their knowledge back to the archive... and...draw upon the wisdom of the thousands of interested researchers, scholars, and students who work with these materials" (Leon 2014). This combination of metadata enrichment and user engagement is the defining feature of crowdsourced humanities transcription platforms. Natural history transcriptions can benefit from an additional layer that draws from the record-based implementation of transcription tools to generate structured datasets, which is most effectively accomplished via markup.

Scripto is available as a plugin for common content management systems (CMS) including Omeka, Wordpress, and Drupal, and best serves projects that use a CMS as a repository for their project content. This model, adopted by the University of Iowa and Wellesley College creates a platform outside of the digital library interface to which images of the digitized items are uploaded and attached to a digital item record that includes a transcription file. Transcriptions are typed by users into a text box with some style convention and slated for verification upon completion. In addition to transcribing activities, volunteers can communicate about items via Disqus[10] annotations or a discussion form that facilitates communication between transcribers and administrators.

The University of Iowa's first transcription project, Civil War Diaries and Letters[11], was launched in 2011 and successfully enhanced collection access and usability by "enabling full-text search of the content, and engaging the general public by allowing them to interact with the materials in new ways" (Saylor and Wolfe 2011). After the popularity of the program crashed their servers, UI sought a more user friendly and efficient solution in RRCHNM's customizable Scripto. The transcriptions of Anne Whitney's correspondence collection[12] was developed out of a Wellesley College undergraduate seminar to provide the Wellesley community witha lifetime learning resource and enable access to the documents for a wider audience (Bartle 2014). Transcribed text allowed Prof. Jacqueline Musacchio to use a range of DH tools to visualize Whitney's travel experience as clearly as possible by capturing evidence of movement through space and time and applying that evidence to historical maps (Musacchio 2014). This is a critical similarity between humanities research and natural history research that transcriptions support. By providing access to transcribed text that connects to specimen occurrences with specific dates and geographic locations, natural historians can identify a broader and more complete picture of biodiversity.

As part of the development process for Scripto, RRCHNM drew from other successful crowdsourcing programs including University College London's (UCL) Transcribe Bentham.

[7] http://scripto.org/.
[8] http://blogs.ucl.ac.uk/transcribe-bentham/.
[9] http://wardepartmentpapers.org/.
[10] https://disqus.com/.
[11] https://diyhistory.lib.uiowa.edu/collections/show/8.
[12] http://omeka.wellesley.edu/whitneytranscribe/home.

Archivists and researchers at UCL identified the need for a fully transcribed facsimile of Jeremy Bentham's papers in order to better facilitate research and eventually publish a complete scholarly edition of Bentham's collected works for wider dissemination.

The Transcribe Bentham digital platform, known as the "Transcription Desk," is a MediaWiki customization that supports TEI encoding via a toolbar for ease of use. While mark-up was not required from volunteers, its adoption and use was significant and demonstrated that "volunteer labor can be used to undertake the type of detailed...tasks generally perceived to be the preserve of those trained in XML (eXtensible Markup Language) and TEI" (Causer and Terras 2014). The validated transcriptions produced by Transcribe Bentham were uploaded to UCL's digital repository and linked to the relevant manuscript item, enhancing access to the collection and improving primary source research. The UCL project was among the first large scale crowdsourcing projects for transcribing special collections items, and reflected a new focus in digital humanities scholarship: increasing and encouraging user engagement and providing open source tools that can be repurposed for other projects (Causer and Terras 2014). Transcribe Bentham was instrumental in demonstrating the sustainability of crowdsourcing approaches to both humanities scholarship and enhancing access to content in libraries in special collections.

SMITHSONIAN TRANSCRIPTION CENTER

The Smithsonian's Transcription Center[13], built in 2012, is a popular system that is designed to extract full text transcriptions from archival collections. Smithsonian Libraries created a flexible program for 19 museums and archives that includes transcription, translation, and discussion features. To accommodate different formats, the team developed several different data structures for field notebooks, diaries, botanical specimen records, and numismatic proofs. The Transcription Center generates JSON (JavaScript Object Notation) files from text entered into a single data field. Volunteers can utilize a WYSIWYG-like toolbar that applies some TEI-compliant markup but minimizes UI interference with the actual process of transcribing. The JSON-stored data allows any type of data to be stored in one database field instead of across several specific tables and can fairly easily interact with XML systems (Gunther, Schall, and Wang 2016).

One of the most significant impacts from the Transcription Center has been its contribution tounderstanding and leveraging motivations of theirvolunteers. Dr. Meghan Ferriter, a former platform coordinator, has written and been interviewed extensively about the value of understanding volunteer motivations and customizing crowdsourcing activities to best address them (Decker 2016; Parilla and Ferriter 2016; Floyd 2017; Ashenfelder 2016; Ferriter 2016, 2014). The Smithsonian outreach and engagement strategies are essential to a project's success and must include communicating in a sincere and authentic way, volunteering information and content, and asking for help. The Biodiversity Heritage Library has strong foundation of cooperative dialogue with its users and would likely be successful in managing a transcription project in a similar way. Learning from the contributions and experience of the Smithsonian Transcription Center offers valuable insight into data management and volunteer engagement practices. The Transcription Center, however, exclusively serves Smithsonian institutions, which is incompatible with the larger organizational structure of BHL.

THE ZOONIVERSE

The largest crowdsourcing program, the Zooniverse[14], redesigned the traditional model for document-based transcription projects. The Zooniverse draws from a long legacy of citizen

[13] https://transcription.si.edu/.

[14] https://www.zooniverse.org.

science and co-creation projects from across scientific disciplines. What began as a web application to invite members of the public to classify and describe the universe's photographed galaxies[15] turned into the world's largest platform for "people powered research". The Zooniverse has largely been focused on digital datasets that require hundreds of thousands if not millions of hours to investigate and classify by leveraging human intuitions and pattern finding abilities. With transcriptions, the image of a handwritten document becomes digital data and the Zooniverse can use a similar model to apply and extract classifications, metadata structures, and bodies of text from manuscript collections.

The Zooniverse's Project Builder and Scribe (Beaudoin 2015) utility that was developed in partnership with the New York Public Library both rely on the concept of microtasking to break up labor intensive transcriptions that require high levels of intelligence and concentration. By splitting tasks into more manageable chunks with varying degrees of difficulty, citizen scientists can engage with the project at whatever level they desire. Breaking up the tasks also improves the data quality by mitigating against user fatigue and boredom. In order to achieve varying project goals successfully, Zooniverse has developed three similar systems that each combine some degree of a Mark, Transcribe, and Verify workflow (Zooniverse 2015; Snakeweight 2016; Simpson, Page, and Roure 2014).

Instead of inviting volunteers to type complete page transcriptions into a text box, they break up the process into three separate tasks. Page words and lines must first be Marked by users to identify text locations in the image to maintain the author's explicit layout and formatting choices; Marked sections are then Transcribed, which transforms the information into machine readable content and preserves the relationship between pixels and text; and Transcribed text must finally be verified for quality control. Output data can be harvested raw (from each task) or algorithmically aggregated (from the whole set of Mark and Transcribe tasks for a given image) along with the level of the Zooniverse's confidence in the accuracy of the transcription. The output data is structured similarly to the Transcription Center's (JSON), but does not depend on a specific database management system to manage data between the transcription platform and the collection repository.

FROMTHEPAGE

FromThePage[16] is a lightweight, open source, collaborative transcription platform. It's defining feature is its use of wiki style markup to link references and subjects within texts to dynamically index terms. The design is optimized for archives projects, is a simple tool and can be deployed quickly. It has a very clean interface for viewing, transcribing, and coding people, places, and subjects across a collection of documents (Figure 7). Since 2005, FromThePage has hosted projects across the humanities and natural sciences and its creators Ben and Sara Brumfield have become important community members that write, blog, and present extensively on their work and the larger themes or trends within the field of manuscript transcription and collaborative digitization. Indeed, FromThePage has built a very good reputation among librarians, archivists, and other project managers over the last decade for its creative implementation of the wiki-like annotation index (Lawson 2012).

Several natural history and botanic institutions have selected FromThePage to transcribe archival manuscript content including the San Diego Museum of Natural History, University of California, Berkeley's Museum of Vertebrate Zoology, and recently, the New York Botanical Garden's LuEsther T. Mertz Library. Each have identified the necessity of transcribing field notes to support research in the natural sciences by locating information about the

[15] https://www.galaxyzoo.org/.

[16] https://fromthepage.com/.

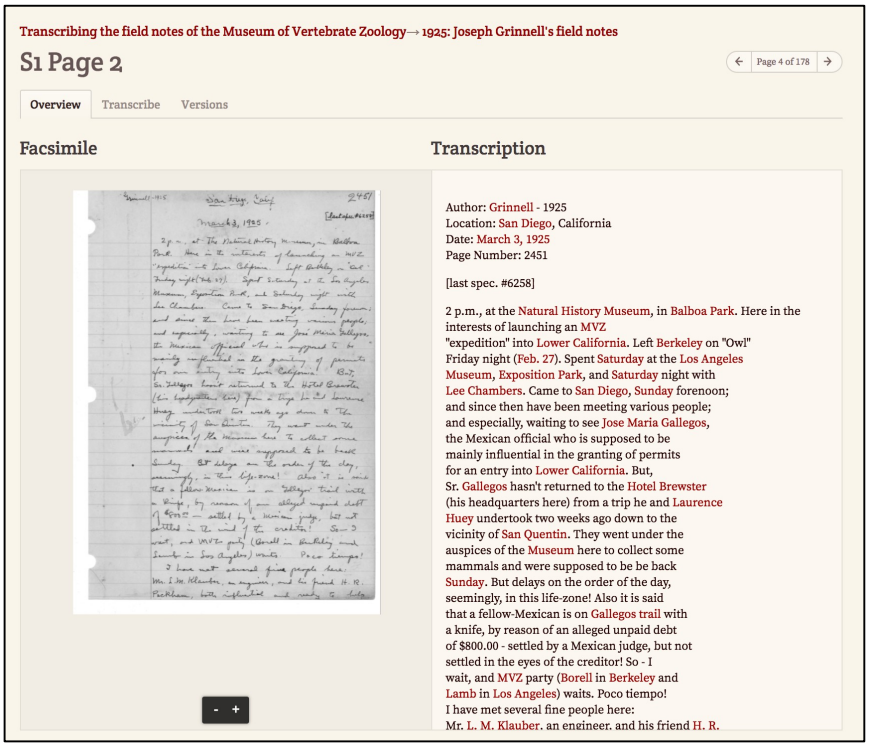

Figure 7. Image of the University of California, Berkeley's Museum of Vertebrate Zoology collection of Joseph Grinnell's field notes that displays the item's image and transcription with MediaWiki encoded metadata tags for people, locations, dates, organizations, and could potentially be used to tag common and Latin scientific names. https://www.fromthepage.com/display/display_page?page_id=4254.

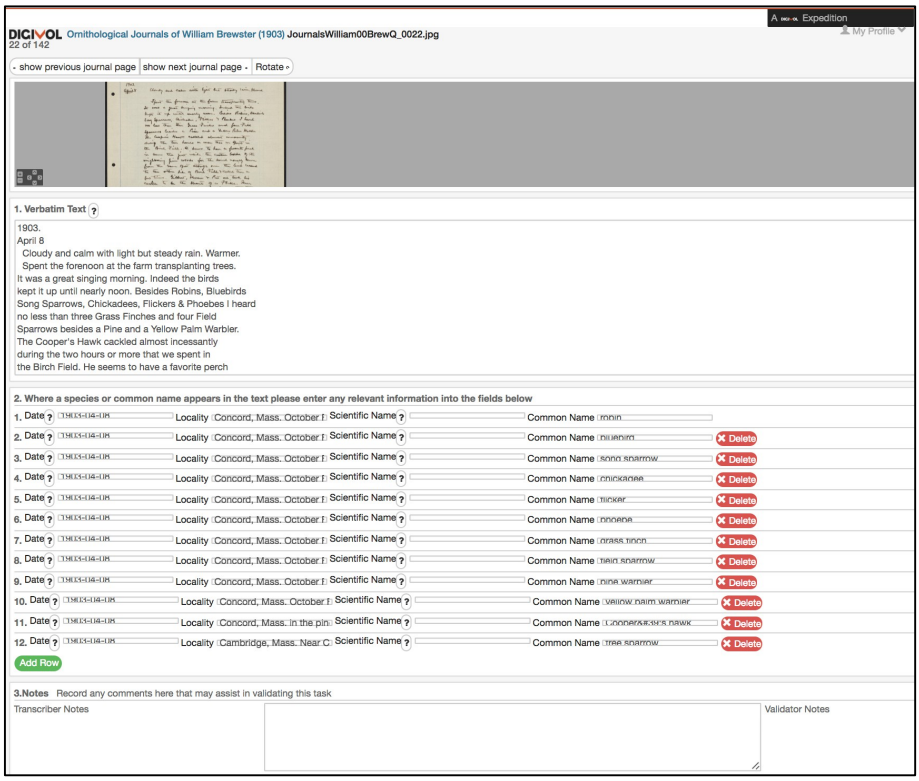

Figure 8. Image of the Museum of Comparative Zoology's collection of ornithologist William Brewster's DigiVol transcription project that displays the platform's ability to add scientific names identified in the image as structured data. http://volunteer.ala.org.au/task/show/17794890.

quantity of species in an area, how often they were observed, and physical attributes of specimen. The San Diego Museum of Natural History has digitized the field notes of noted herpetologist Laurence M. Klauber and commenced their transcription project in 2010.

By March 2014 10,000 subjects had been identified for classification and volunteers made 24,000 page edits and 42,000 links between individual observations, species names, and personal names (Brumfield 2014). Similarly, the Museum of Vertebrate Zoology at Berkeley transcribed field notes of Joseph Grinnell and preserve his unique method of recording field observations. Grinnell and other MVZ scientists recorded observation notes with a particular philosophy (Grinnell 1910) and using a precise structure that does not vary across creators, and, therefore, FromThePage's tagging and indexing feature can be leveraged with tremendous success to create structured datasets for the collections. The New York Botanical Garden is also in the first steps of developing a transcription project for the John Torrey Papers that includes correspondence, manuscripts, notes, and botanical illustrations.

DIGIVOL

DigiVol[17] was built by the Australian Museum as an Atlas of Living Australia project and is designed primarily for transcription of biodiversity-related materials (Kearney and Wallis 2015). It combines a simple and attractive viewing and transcription interface with tools for extracting specimen data from items. There is no easy process for marking up text, but the platform features a form that invites volunteers to enter scientific names of specimens with dates and locations of their collection or observation. This generates a CSV document that retains valuable information in a structured format.

DigiVol is available for use by any institution, and simply requires establishment of a free account. Project managers are given administrative privileges that enable uploading of content, creation of custom guidelines and tutorials for each project, review and validation of the work of transcribers, and export of transcription files in CSV format.

DigiVol is nicely designed to facilitate communication between volunteers and project administrators. Each transcription page has a comment box so that volunteers can make notes or ask questions about that page. There is a forum on which transcribers can raise topics and post comments or questions about any field notebook, and these are visible to administrators as well as other volunteers. Managers can proactively use the forum to call attention to challenging pages, or address commonly made errors. Volunteers can also email project administrators directly with queries or suggestions.

The Ernst Mayr Library of the Museum of Comparative Zoology has been using DigiVol since 2014 for transcription of the field notes of William Brewster, an amateur ornithologist of the late 19th and early 20th centuries. Both DigiVol and FromThePage were initially used by the Library in 2014-2015 to transcribe Brewster materials for the Purposeful Gaming project. To date, over 6300 pages have been transcribed via DigiVol, and the resulting exports will form part of the initial load of transcription files to BHL. Species names and other structured occurrence data, captured apart from the transcriptions and exported as a separate CSV file, will be retained for possible future use in BHL pending development of a new data model (Figure 8).

Early transcription platforms were generally tied to specific manuscript collections of significant value or at high risk for long term preservation. These projects depended on curation and were optimized for smaller programs for which significant manual data and community management practices were plausible. Workflows for these projects also often require significant manual processes including the selection and curation of collections, potentially digitizing content if

[17] http://volunteer.ala.org.au/.

necessary, uploading image files to external transcription platform, conducting outreach activities to recruit volunteers, writing collections specific tutorials, answering questions regarding transcriptions and data entry processes, validating completed works and performing QA, exporting text and associated metadata, transforming text and data into schema and formats optimal for a specific digital repository, uploading transcriptions to repository, ensuring that access is appropriate, and engaging in general troubleshooting.

These models do not typically scale well, even with the expertise of a dedicated community or program manager. Outputs for document-based projects are largely simple plain text documents that improve readability (by humans and machines) and do not structure data outputs. Conversely, record-based programs enrich collections with structured data and metadata, but are limited in their capacity to accurately represent works in their entirety. The time investment required for collections-based transcription programs and their technological limitations prohibit or limit their use for large digital collections like the Biodiversity Heritage Library. As the field of manuscript transcriptions developed and crowdsourcing proved to be a reliable method for generating machine readable text, platforms evolved to meet the demands of larger digital collections. The Zooniverse and DigiVol incorporated citizen science practices and ideas and FromThePage and the Smithsonian Transcription Center included markup and tagging features to structure text and data to improve access and enrich collections metadata.

DISCUSSION

The web is a co-creative digital experience, and GLAM organizations need to be prepared to engage with users' knowledge and experience to build and augment online content (Terras 2016). Successful crowdsourcing platforms do more than invite users to donate time and expertise to specific projects; they cultivate digital environments that encourage open access and the clear and open transfer of ideas. In digital libraries, collections are not hidden behind glass exhibition cases but are living texts and documents that operate as part of a wider ecosystem of knowledge sharing and co-creation.

BHL plans to design a workflow for its partners to generate high quality transcriptions which can be deployed with relative ease. Field notebook and manuscript transcriptions need to fit into the larger BHL objective of producing and making available large scale datasets for users and researchers to study and manipulate. While generating only full-text transcriptions will improve discoverability, manuscript items will remain isolated from the other literature collections. One way to link knowledge produced in field notes and information derived from books and journals in BHL is to extend the database by adding tables and relationships for common access points including common species names, locations, dates, identified people and organizations, and events (Figure 6) (Studer and Rinaldo 2014). Item level catalog records for monographs and serials contain publication information that enhances access and provides context. Special collections content, however, often does not include basic item level contextual information like creator (author), creation (publication) date, subject headings, location/geographic information, or language.

As it designs a transcription and data collection program, BHL can leverage lessons learned from established crowdsourcing projects in the cultural heritage sector as well as citizen science initiatives that are common in biodiversity and natural history domains (Table 1). As a digital library focused on scientific inquiry, BHL is well situated to join the "Collections as Data" movement by developing programs, including transcription and data collection, to support computational analyses and distant reading of texts (Zwaard 2017). In order to truly capitalize on the power of digitization, future versions of BHL will aim to support text and data mining of its content and collections metadata.

Table 1: Comparison of transcription tools treated in the text.

Platform	Projects	Advantages	Disadvantages
Scripto[1]	Papers of the War Department[2]; Anne Whitney papers[3]	Can be integrated with content management systems, including Omeka, Wordpress, and Drupal plugins	Collection-based; does not scale up particularly well
Transcribe Bentham's Transcription Desk[4]	Transcribe Bentham[5]	Designed as a research project; reveals much about project design	Collection based; requires significant amount of customization and development
Smithsonian Transcription Center[6]	William M. Mann Field Notes - Fiji and British Solomon Islands, 1915-1916[7]; Albert Spear Hitchcock Field Notes[8]	Extremely successful institutional program with significant number of volunteers; potential for connecting data from special collections and archives with Smithsonian specimen collections	Only available for use by Smithsonian Institution organizations
Zooniverse[9]	Anno.Tate[10] Shakespeare's World[11]; Beyond Words[12]	Largest volunteer base; deeply connected to the citizen science universe; scales up well	Mark and Transcribe workflow interrupts user interface connection to document; aggregation method requires multiple volunteers to complete tasks; does not provide immediate user access to output
FromThePage[13]	Joseph Grinnell's field notes[14]; C.S. Pierce manuscripts[15]	MediaWiki tags are flexible and simple to use; cleanest user interface; flexible; open source development; supports OCR corrections in addition to manuscript transcriptions	External platform requires some development to improve interoperability
DigiVol[16]	William Brewster ornithological journals[17]	Includes optional forms for entering structured occurrence data optimized for natural history collections	Occurrence data must be entered *in addition to* transcription

[1] http://scripto.org/.
[2] http://wardepartmentpapers.org/.
[3] http://omeka.wellesley.edu/whitneytranscribe/home.
[4] http://www.transcribe-bentham.da.ulcc.ac.uk/td/Transcribe_Bentham.
[5] http://www.ucl.ac.uk/transcribe-bentham.
[6] https://transcription.si.edu/.
[7] https://transcription.si.edu/project/9708.
[8] https://transcription.si.edu/project/10894.
[9] https://www.zooniverse.org/.
[10] https://anno.tate.org.uk.
[11] https://www.shakespearesworld.org.
[12] http://beyondwords.labs.loc.gov.
[13] https://fromthepage.com/.
[14] https://fromthepage.com/cfidler/transcribing-the-field-notes-of-the-museum-of-vertebrate-zoology.
[15] https://fromthepage.com/jeffdown1/c-s-peirce-manuscripts.
[16] https://digivol.ala.org.au/.
[17] https://digivol.ala.org.au/institution/index/11740375.

Digitizing items in archives and special collections and adding them to online repositories promised dramatically enhanced access but did not deliver. Images of content are still described at collection levels and are perhaps not as useable as once imagined. Transcribing textual information contained in these images facilitates indexing and searching. Texts can be mined to enrich metadata attributes, and context can be applied to records to better connect items according to content. Transcriptions, however, are also in danger of disappearing into digital repositories. Without some kind of imposed intellectual framework, digitized items are lost to the "dank cellar of electronic texts" (Shillingsburg 2006), which BHL is working to avoid by developing crowdsourcing initiatives in concert with redesigning the portal's metadata framework, image delivery system, and taxonomic backbone.

LITERATURE CITED

Ashenfelder, M. 2016. 'Volun-peers' Help Liberate Smithsonian Digital Collections. The Signal Blog. December.[18]

Bartle, J. 2014. The Letters of Anne Whitney: Using Archives in Digital Scholarship. Feminist Collections 35 (3/4).

Beaudoin, P. 2015. Scribe: Toward a General Framework for Community Transcription. NYPL Labs. November.[19]

Blaser, L. 2014. Old Weather: Approaching Collections from a Different Angle. In Crowdsourcing Our Cultural Heritage, edited by Mia Ridge, 45–55. Ashgate Publishing Ltd., Surrey.

Bonney, R, C.B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K.V. Rosenberg, and J. Shirk. 2009. Citizen science: A developing tool for expanding science knowledge and scientific literacy. BioScience 11:977–984. doi:10.1525/bio.2009.59.11.9.

Brabham, D.C. 2013. Crowdsourcing. MIT Press, Cambridge.

Brewster, W. 1892. Journals of William Brewster, 1871-1919. Museum of Comparative Zoology.[20]

Brumfield, B. 2012. What does it mean to 'support TEI' for manuscript transcription? Collaborative Manuscript Transcription. November 10.[21]

Brumfield, B. 2014. Wikilinks in FromThePage. Collaborative Manuscript Transcription. March 14.[22]

Causer, T., and M. Terras. 2014. Crowdsourcing Bentham: Beyond the traditional boundaries of academic history. International Journal of Humanities and Arts Computing 8:46–64. doi:10.3366/ijhac.2014.0119.

Decker, J. 2016. Exploring the Smithsonian Institution Transcription Center [Special Issue]. Collections: A Journal for Museum and Archives Professionals 12(2). Rowman/Littlefield.

DeVeer, J.M., C.A. Rinaldo, and L. Ford. 2013. Primary source material in science: the importance of archival field notes.[23]

Dickinson, J.L., and R. Bonney, eds. 2012. Citizen Science: Public Participation in Environmental Research. Cornell University Press, Ithaca.

Dickinson, J.L., J. Shirk, D. Bonter, R. Bonney, R.L. Crain, J. Martin, T. Phillips, and K. Purcell. 2012. The current state of citizen science as a tool for ecological research and public engagement. Frontiers in Ecology and the Environment 10(6). doi:10.1890/110236.

Ellis, S. 2014. A history of collaboration, a future in crowdsourcing: positive impact of cooperation on British librarianship. Libri 64(1):1–10. doi:10.1515/libri-2014-0001.

Europe, Council of. 2005. Council of Europe Framework Convention on the Value of Cultural Heritage for Society. Council of Europe Treaty Series No. 199.

Evans, M. 2007. Archives of the people, by the people, for the people. American Archivist 70(2):387–400. doi:10.17723/aarc.70.2.d157t6667g54536g.

Ferriter, M. 2014. Growing to a Community of Volunpeers: Communication and Discovery. Smithsonian Institution Archives. July.[24]

[18] http://blogs.loc.gov/thesignal/2016/12/volun-peers-help-liberate-smithsonian-digital-collections/.
[19] https://www.nypl.org/blog/2015/11/23/scribe-framework-community-transcription.
[20] https://www.biodiversitylibrary.org/bibliography/77525.
[21] http://manuscripttranscription.blogspot.com/2012/11/what-does-it-mean-to-support-tei-for.html.
[22] https://manuscripttranscription.blogspot.com/2014/03/.
[23] http://escholarship.umassmed.edu/esciencesymposium/2013/posters/14.
[24] https://siarchives.si.edu/blog/growing-community-volunpeers-communication-discovery.

Ferriter, M. 2016. Volunpeers: Hashtag, Identity, and Collaborative Engagement. Meghan in Motion Blog. April.[25]

Floyd, S. 2017. Engaging with volunteers: Smithsonian Transcription Center. FromThePage Blog, April 2017.[26]

Grinnell, J. 1910. The methods and uses of a research museum. Popular Science Monthly 77:163–169.

Gunther, A, M. Schall, and C.-H. Wang. 2016. The creation and evolution of the Transcription Center: Smithsonian Institution's digital volunteer platform. Collections 12(2): 87–96.

Gwinn, N.E., and C.A. Rinaldo. 2009. The Biodiversity Heritage Library: Sharing biodiversity literature with the world. IFLA Journal 35(1):25–34.

Hanse, D., J. Gehring, P. Schone, and M. Reid. 2012. Improving Indexing Efficiency and Quality: Comparing A-B-Arbitrate and Peer Review. Family History Technology Workshop, Brigham Young University, Provo.

Holley, R. 2010. Crowdsourcing: How and why libraries should do it. D-Lib Magazine 16 (3/4). doi: 10.1045/march2010-holley.

Kalfatovic, M., and C. Rinaldo. 2016. Enabling Progress in Global Biodiversity Research: The Biodiversity Heritage Library. In Libraries: Enabling Progress, Proceedings of the Eighth Shanghai International Library Forum. Shanghai Scientific / Technological Literature Press. Pp. 406–418.

Kearney, N., and E. Wallis. 2015. Transcribing between the lines: crowd-sourcing historic data collection. MWA2015: Museums and the Web Asia 2015.[27]

Lawson, K.L. 2012. Crowdsourcing transcription: FromThePage and Scripto. Chronicle of Higher Education.[28]

Leon, S.M. 2014. Build, analyze, and generalize: Community transcription of the papers of the War Department and the development of Scripto. Pp. 89-111 in Crowdsourcing Our Cultural Heritage (M. Ridge, ed.). Ashgate Publishing Ltd., Surrey.

McClanahan, P. 2017. Getting to Know the BHL Users. May.[29]

Miller-Rushing, A., R. Primack, and R. Bonney. 2012. The history of public participation in ecological research. Pp. 285-290 in Citizen Science: Public Participation in Environmental Research (J.L. Dickinson and R. Bonney, eds.). Cornell University Press, Ithaca. doi:10.1890/110278.

Musacchio, J.M. 2014. Project Narrative: Anne Whitney Abroad, 1867-1868.[30]

Parilla, L., and M. Ferriter. 2016. Social media and crowdsourced transcription of historical materials at the Smithsonian Institution: Methods for strengthening community engagement and its tie to transcription output. American Archivist 79(2). doi:10.17723/0360-9081-79.2.438.

Pilsk, S.C., M.A. Person, J.M. DeVeer, J.F. Furfey, and M.R. Kalfatovic. 2010. The Biodiversity Heritage Library: Advancing metadata practices in a collaborative digital library. Journal of Library Metadata10: 136–155. doi:10.1080/19386389.2010.506400.

Rinaldo, C.A., and J. Smith. 2014. Moving through time and culture with the Biodiversity Heritage Library. Pp. 95-108 in Migrating Heritage: Experiences of Cultural Networks and Cultural Dialogue in Europe (P. Innocenti, ed.). Ashgate Publishing Ltd., Surrey.

Rose-Sandler, T. 2012. The Art of Life: Data Mining and Crowdsourcing the Identification and Description of Natural History Illustrations from the Biodiversity Heritage Library. Biodiversity Heritage Library.[31]

Rose-Sandler, T. 2015. Purposeful gaming and BHL: engaging the public in improving and enhancing access to digital texts. Biodiversity Heritage Library.[32]

Saylor, N., and J. Wolfe. 2011. Experimenting with Strategies for Crowdsourcing Manuscript Transcription. Research Library Issues: A Quarterly Report from ARL, CNI, and SPARC.[33]

Schmitt, W.L. 1962. Palmer Peninsula (Antarctica) Survey, 1962-1963: miscellaneous notes (2 of 4). Series: SIA RU007231.[34]

Shillingsburg, P. 2006. From Gutenberg to Google: Electronic Representations of Literary Texts. Cambridge University Press, Cambridge.

[25] http://meghaninmotion.com/2016/04/05/volunpeers-hashtag-identity-engagement/.
[26] http://content.fromthepage.com/smithsonian_volunpeers/.
[27] http://mwa2015.museumsandtheweb.com/paper/transcribing-between-the-lines-crowd-sourcing-historic-data-collection/.
[28] http://www.chronicle.com/blogs/profhacker/crowdsourcing-transcription-fromthepage-and-scripto/38028.
[29] https://ndsrbhl.wordpress.com/2017/05/17/getting-to-know-the-bhl-users/.
[30] http://www.w2ww.19thc-artworldwide.org/index.php/autumn14/musacchio-project-narrative.
[31] http://biodivlib.wikispaces.com/Art+of+Life.
[32] http://biodivlib.wikispaces.com/Purposeful+Gaming.
[33] https://eric.ed.gov/?q=ED527715&id=ED527715.
[34] https://www.biodiversitylibrary.org/bibliography/130464#/summary.

Silberglied, R.E. 1965. Field Notes, Mexico, July-August 1965. SIA RU007316.[35]

Simpson, R., K.R. Page, and D. De Roure. 2014. Zooniverse: Observing the World's Largest Citizen Science Platform. Pp. 1049-1054 in Proceedings of the 23rd International Conference on World Wide Web. Association for Computing Machinery. doi:10.1145/2567948.2579215.

Snakeweight. 2016. 'What's up with those grey dots?' you ask. Shakespeare's World blog. February.[36]

Studer, M., and C.A. Rinaldo. 2014. From Historical Field Notes to Mobile Field Guides: The Encyclopedia of Life and the Biodiversity Heritage Library Team up to Connect Biodiversity-Related Content across the Centuries to Support Today's Ecological Research and Education Needs. Ecological Society of America.[37]

Terras, M. 2016. Crowdsourcing in the digital humanities. Pp. 420-439 in A New Companion to Digital Humanities. Wiley-Blackwell, Chichester.

Theobald, E.J., A.K. Ettinger, H.K. Burgess, L.B. DeBey, N.R. Schmidt, H.E. Froehlich, C. Wagner, J. HilleRisLambers, J. Tewksbury, M.A. Harsch, and J.K. Parrish. 2015. Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation 181:236–244. doi: 10.1016/j.biocon.2014.10.021.

Zooniverse. 2015. One Line at a Time: A New Approach to Transcription and Art History. Zooniverse. September.[38]

Zwaard, K. 2017. Collections as Data and National Digital Initiatives. Library of Congress. August. [39]

[35] https://www.biodiversitylibrary.org/bibliography/95422#/summary.
[36] https://blog.shakespearesworld.org/2016/02/24/whats-up-with-those-grey-dots-you-ask/.
[37] https://eco.confex.com/eco/2014/webprogram/Paper47651.html.
[38] https://blog.zooniverse.org/2015/09/01/one-line-at-a-time-a-new-approach-to-transcription-and-art-history/.
[39] https://blogs.loc.gov/thesignal/2017/08/collections-as-data-and-national-digital-initiatives/.