Collections as Data:
Crowdsourcing Data Corrections and Enhancements in the Biodiversity Heritage Library

---

BHL NDSR Best Practices and Recommendations Report

Katie Mika
January 8, 2017

As digital libraries scale up, access is no longer defined as "putting things online." Content needs to be situated within a digital landscape that gives it meaning and utility. Archivists understand the concept of "context" as a vital part of developing a comprehensive intellectual arrangement and description of a collection. This framework can be extended to understand the role of library materials in the grander scheme of digital access and use for a knowledge domain. The collective charge for the BHL NDSR Residents is to plan a "next generation digital library with tools based on user priorities to test the efficacy of incorporating transcriptions, image searching, analyzing complex collections, and identifying effective connections that can be made among museums, archives, and biodiversity databases."[1] And "Foundations to Actions: Extending Innovations in Digital Libraries in Partnership with NDSR Learners" identifies the Crowdsourced Data Corrections project goal for the MCZ Resident as "develop[ing] methodologies and propose[ing] tools for the integration of crowdsourced data corrections and enhancements back into the BHL portal."[2] In addition to analyzing the development of crowdsourcing programs in digital libraries for manuscript transcription and Optical Character Recognition (OCR) output corrections, I also researched named entity recognition, large scale digital library data models, Linked Open Data (LOD) systems, and Collections as Data initiatives in order to to develop Best Practices recommendations and understand how text data can be leveraged to better connect BHL content to an emerging biodiversity knowledge graph to support the use and reuse of its heritage literature and manuscript collections.

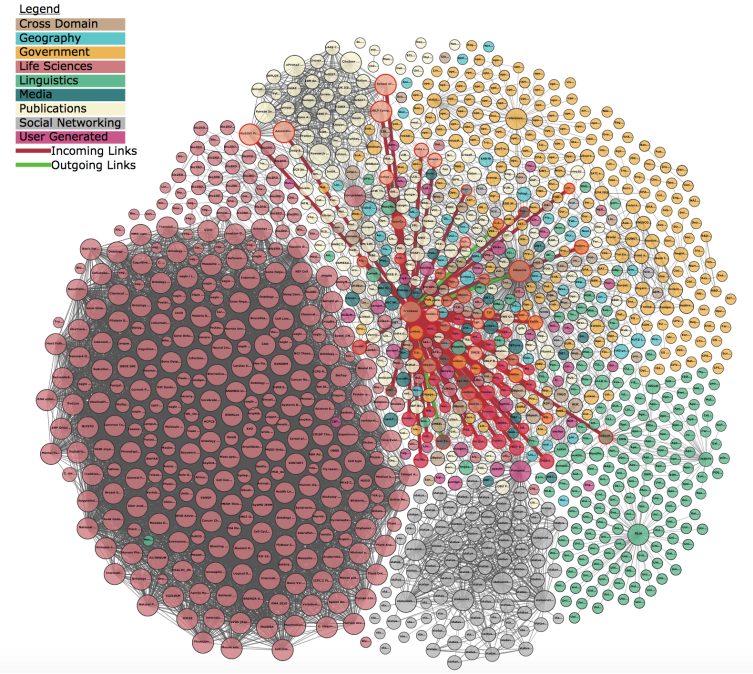
A knowledge graph, simply and in a context relevant to this report, is a cloud of entities with defined relationships that can be pulled and moved around to connect these entities in different ways. Articles, journals, monographs, authors, and institutions are all entities that are given Uniform Resource Identifiers (URIs), which reduce ambiguity, especially in cases where there are several ways to represent the same thing.[3] These entities become nodes on a graph with edges defined by ontologies and controlled vocabularies that describe relationships [Fig 1].

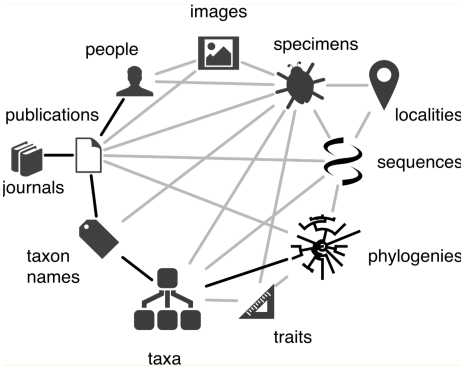
[1] "Foundations to Actions: Extending Innovations in Digital Libraries in Partnership with NDSR Learners." Institute for Museum and Library Services, Laura Bush 21st Century Librarian Program. Award number: RE-40-16-0082-16. https://www.imls.gov/grants/awarded/RE-40-16-0082-16

[2] *Ibid.*

[3] There are some important complexities to consider when identifying entities and assigning URIs, but such a discussion is outside the scope of this paper and project.

Figure 1: The Linked Open Data Cloud Diagram[4] and Biodiversity Knowledge Graph[5]





While there is not a comprehensive knowledge graph for biodiversity (yet), there are several “calls to action” that identify its necessity, the ripe opportunity, and technological capability for its development. Rod Page advocates for this these types of connections [fig. 2] in articles[6], blog posts[7], and conference keynotes[8], and Holly Bik argues in “Let’s rise up and unite taxonomy and technology” that “by tapping the traditional taxonomic knowledge base, we will fundamentally improve database resources for all scientific disciplines, dramatically improving the accessibility of historical knowledge.”[9] Her visualization [Fig. 2] of integrating genomic information with digitized specimen records and morphological taxonomy illustrates the value in making clear connections between disparate datasets across multiple sources. Libraries should aim to include their collections in this burgeoning knowledge graph.

---

[4] http://lod-cloud.net/

[5] Page, Roderic D. M. “Towards a Biodiversity Knowledge Graph.” Conference Keynote Presentation, Biodiversity Information Standards (TDWG), October 2, 2017. http://idigbio.adobeconnect.com/p2ul6gfd20os/

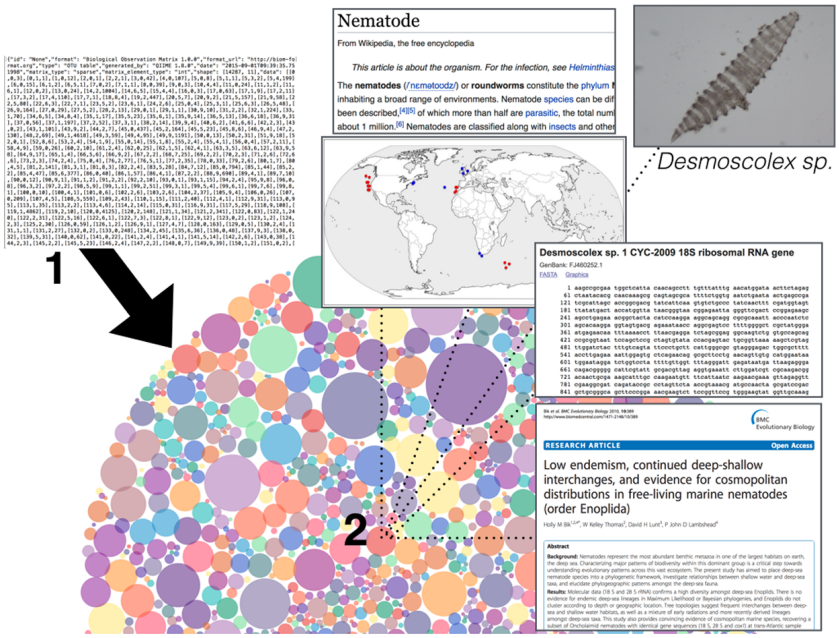
[6] Page, Roderic D. M. “BioNames: Linking Taxonomy, Texts, and Trees.” *PeerJ* 1 (October 29, 2013): e190. https://doi.org/10.7717/peerj.190.

[7] http://iphylo.blogspot.ca/search/label/Linked%20data

[8] Page, Roderic D. M. “Towards a Biodiversity Knowledge Graph.”

[9] Bik, Holly M. “Let’s rise up to unite taxonomy and technology.” *PLOS Biology*. 15(8): e2002231. August 18, 2017. https://doi.org/10.1371/journal.pbio.2002231.

Figure 2: Uniting Taxonomy and Technology[10]

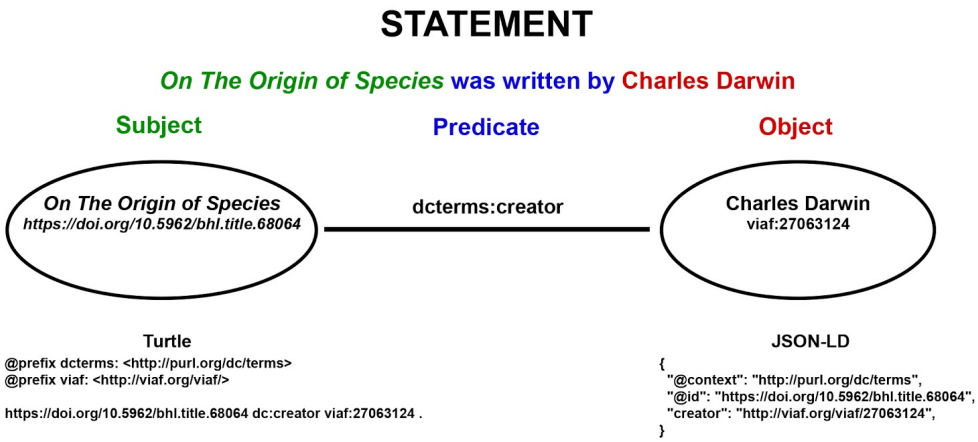


Linked Open Data technologies are being adopted by libraries in order to connect collections content and metadata to other nodes in the LOD cloud. The Semantic Web[11] refers generally to the idea of a web of data (as opposed to today's web of documents), and enables the creation of data stores on the web, the building vocabularies to describe the data, and setting of rules for handling it.[12] The Semantic Web, as it is now defined, is open and relies on linked data to interpret an infinite amount of data resources. Unlike XML, which defines records according to attribute/value pairs, linked data creates individual statements that can be aggregated to create a more complete picture of a resource. Languages like RDF and SPARQL facilitate data modeling and querying, and data models like the Europeana Data Model and BIBFRAME rely on vocabularies or ontologies including SKOS, schema.org, and Dublin Core to describe library objects and their "aboutness." The Europeana Data Model leverages linked data in order to bring together metadata from digitized objects at institutions across Europe.[13] The BIBFRAME (BIBliographic FRAMEwork) data model developed by the Library of Congress aims to replace MARC cataloging standards and use linked data principles to make bibliographic metadata more interoperable within and outside the library community.[14] The Linked Data for Libraries (LD4L) Labs[15] and Linked Data for Production (LD4P)[16] are two collaborative projects

[10] *Ibid.*
[11] "Why Linked Data?" Linked Data for Libraries. https://www.ld4l.org/linked-data
[12] Semantic Web. World Wide Web Consortium. https://www.w3.org/standards/semanticweb/
[13] Europeana Data Model Documentation. Europeana Pro. https://pro.europeana.eu/resources/standardization-tools/edm-documentation
[14] "Bibliographic Framework Initiative." Library of Congress. http://www.loc.gov/bibframe/
[15] Linked Data for Libraries Labe. https://www.ld4l.org/ld4l-labs

designed to document realistic linked data use cases and implementation scenarios in research libraries. In all models, Linked Data triples are created from bibliographic metadata and relationships are defined by ontologies or vocabularies in a Subject -- Predicate -- Object format [Fig. 3].

Figure 3: Linked Data Triple Example



While this is not a project about linked data, it is important to note that BHL already creates and stores a number of entities, including items, authors, pages, and scientific names, and creates URIs [Fig. 4].

Figure 4: Table with BHL Entities and their URIs

Entity	URI
The Edible Clams, Mussels, and Scallops of California	https://doi.org/10.5962/bhl.title.10054
Weymouth, Frank Walter	https://www.biodiversitylibrary.org/creator/8571
Edible Clams page image	https://biodiversitylibrary.org/page/10603637

A transition to a linked data graph database would allow BHL to create and store statements about these and other entities that would support queries (via SPARQL endpoint) in applications spanning the biodiversity knowledge graph. As library resources are increasingly added to the

[16] Linked Data for Libraries. Harvard University Libraries Wiki. https://wiki.duraspace.org/pages/viewpage.action?pageId=74515029

Semantic Web of Data, Best Practices in archival science are similarly shifting to an understanding of manuscript and records collections as research data resources.

Arguably beginning (in earnest) with the Computational Turn[17] in digital humanities domains and the increase in the availability of machine readable Big Data for scientific research, GLAM institutions have augmented their support of computational uses of collections items and metadata. The Biodiversity Heritage Library at its inception in 2007 committed to bringing open access to digitized biodiversity materials. By supporting metadata harvesting via OAI-PMH, providing data exports and dumps, and writing APIs to give access to data programmatically BHL continues to embrace digital and computational use and reuse of its collections. Within the fields of Library and Information Science and Archives and Records Management, the development of Collections as Data and Computational Archival Science (CAS) initiatives to create frameworks for understanding GLAM organizations and collections as valuable repositories of research data for all domains.

Collections as Data is being defined in the IMLS grant charter project "Always Already Computational" that brings together partners from academic and public institutions to foster a "strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research and teaching."[18] Principal Investigator Thomas Padilla in partnership with archivists and librarians from the University of Pennsylvania, the University of North Carolina Chapel Hill, Penn State University, Emory University, and Texas A&M University argue that digital collection development largely focuses on replicating or imitating traditional analog methods of interacting with objects in a digital environment and that this approach no longer meets the needs of users who are interested in leveraging computational methods and tools to treat digital library collections as data.[19] The program website identifies 28 partners from academic and public institutions in the United States, United Kingdom, and Australia and 14 facets (projects) with a current open call for a new host in the upcoming year detailing implementation projects to present a "multifaceted argument for the present and future state of collections as data."[20]

Computational Archival Science (CAS) is a similar program spearheaded by the University Of Maryland iSchool's Digital Curation Innovation Center that aims to apply computational theory, principles, and methods to a deeper and more integrated blend of archival theory, principles, and methods in order to support the growing complexity of information systems and the processing and preserving of digital records.[21] While Collections as Data

---

[17] Berry, David M. "The Computational Turn: Thinking About the Digital Humanities." *Culture Machine* 12 (February 18, 2011). http://www.culturemachine.net/index.php/cm/article/view/440.
[18] https://collectionsasdata.github.io/
[19] Thomas Padilla, Laurie Allen, Stewart Varner, Patricia Hswe, Elizabeth Roke, and Sarah Potvin. "Always Already Computational: Library Collections as Data." Institute for Museum and Library Services. National Leadership Grants for Libraries, National Digital Platform Project Proposal. 2016.
[20] https://collectionsasdata.github.io/facets.
[21] Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. "Archival Records and Training in the Age of Big Data." In *Advances in Librarianship* -

focuses on archival strategies to provide access to material in ways and formats amenable to computational research, CAS aims to bring computational methods into the theory and practice of creating and maintaining archives and records collections. Its foundational paper "Archival Records and Training in the Age of Big Data" proposes a working definition of Computational Archival Science:

> A transdisciplinary field concerned with the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access, with the aim of improving efficiency, productivity, and precision in support of appraisal, arrangement and description, preservation, and access decisions.[22]

The December 2017 CAS workshop at the IEEE Big Data 2017 conference featured fourteen conference papers, two demos and, one student session focused on exploring the conjunction (and its consequences) of emerging methods and technologies around big data with archival practice and new forms of analysis and historical, social, scientific, and cultural engagement with archives.[23]

Along with Linked Open Data, these movements demonstrate the growing imperative to facilitate and apply automated processes to machine readable GLAM collections data and metadata. Machine readable facsimiles of BHL collections are most often text files that are extraordinary data sources for a range of computational processes. Several projects within the biodiversity scientific community describe the process of mining BHL texts for research data. The workshop paper "EMODnet Workshop on mechanisms and guidelines to mobilise historical data into biogeographic databases" outlines some of the problems associated with extracting species occurrence data from legacy biodiversity data.[24] TDWG 2017 featured the workshop "Construction of biodiversity knowledge graphs driven by federated text mining tools," which included sessions describing the use of published and grey literature in BHL for data curation including entity extraction, natural language processing, and machine learning processes to identify semantic relationships within content.[25] While full text and faceted searching undoubtedly improves the findability of digitized BHL content, these papers describe imperfect (or downright unusable in the case of handwritten documents) OCR processes as a major

---

Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education. Eds.: Lindsay C. Sarin, Johnna Percell, Paul T. Jaeger, & John Carlo Bertot. In Press. http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/Marciano_Kurtz_et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final-1.pdf

[22] Richard Marciano, *et al*. "Archival Records and Training in the Age of Big Data."

[23] Computational Archival Science Workshop, IEEE Big Data 2017. http://dcicblog.umd.edu/cas/ieee_big_data_2017_cas-workshop/

[24] Faulwetter, Sarah, Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Donat Agosti, Christos Arvanitidis, Laura Boicenco, et al. "EMODnet Workshop on Mechanisms and Guidelines to Mobilise Historical Data into Biogeographic Databases." *Research Ideas and Outcomes* 2 (June 7, 2016): e9774. https://doi.org/10.3897/rio.2.e9774.

[25] "Construction of biodiversity knowledge graphs driven by federated text mining tools." *Biodiversity Information Standards, TDWG 2017*. Ottawa, Ontario, October 2017. https://biss.pensoft.net/browse_user_collection_documents?collection_id=44

impediment for text mining. If BHL truly wants to capitalize on the power of digitization, it should aim to support and improve opportunities for text and data mining of its content.

With over 53 million pages of text and image data BHL is well situated to join the Collections as Data movement. Indeed, BHL's long term goals include mining the vast amount of biodiversity data locked away in field notes, correspondence, images, and published texts and making the data accessible from the portal.[26] Available information in BHL for a given species could then potentially include its original description in the published literature linked to an account of the field collection of the type specimen used as a basis for the description and to images in texts and artwork. Eventually, links to the museum specimen and other non-literature-based data and objects will collocate all the elements of biodiversity information. Transcription and tagging programs can support this work by identifying and encoding content that can be connected to other literature and archives collections. The grant projects Connecting Content and Mining Biodiversity identified the utility in connecting these data and developed some workflows and description ontologies that can be recycled in a transcription program.

Fully titled "Connecting Content: A Collaboration to Link Field Notes to Specimens and Published Literature", the collaboration between the California Academy of Sciences, Missouri Botanical Garden, Academy of Natural Sciences of Philadelphia, Harvard University Botany Libraries, Harvard Museum of Comparative Zoology Library, the New York Botanical Garden, and the Smithsonian National Museum of Natural History aimed to "improve access to biodiversity research materials in archives, libraries, and museums through identifying connections and creating linkages between collections that are also intellectually dispersed in terms of collections management practices."[27] Contributing libraries carried out the intellectual legwork to discover and extract data that can be linked between field books, literature, and specimen collections. This information was used to define a sample set of items and assess the frequency of different types of connections, and a set of recommendations was created for developers to expand functionality to support links between related content. This work however, was carried out manually, which is not terribly sustainable and does not scale well. The primary conclusion from Connecting Content supports the understanding that taxon names are important access points and links between collections. It follows that adding resources that enhance these connections and extending the concept to include entities like geographic locations, added people and corporate bodies, events, and trait descriptions would be similarly beneficial. Another significant take-away included an analysis of BHL metadata.

> There are a lot of inconsistencies and we're trying to look for anything that will yield results to connect these materials but so far, we're not having a ton of luck. This is in

[26] Mika, Katherine A., Joseph De Veer, and Constance Rinaldo. "Crowdsourcing Natural History Archives: Tools for Extracting Transcriptions and Data." *Biodiversity Informatics* 12, no. 0 (November 14, 2017). https://doi.org/10.17161/bi.v12i0.6646.

[27] Fischer, Richard T. "Connecting Content: A Collaboration to Link Field Notes to Specimens and Published Literature." Society of American Archivists. May 2012. https://www2.archivists.org/sites/all/files/RichardFischer-SAA-ResearchPaper-2011.pdf

part because some of the data we went out is not consistent and in part because some of the data we sent out is being parsed out in ways that aren't useful of metadata is buried or discarded for a variety of different reasons.[28]

There are a range of opportunities for metadata enrichment in a digital library environment that BHL may investigate further. Engaging in authority work and data validation and reconciliation between disparate data sources is vital for serving users and managing hundreds of thousands of collection items. As additional entity recognition processes or annotation features are added to the BHL portal, messy metadata will explode exponentially and cause a host of future problems. Currently, BHL creators are being disambiguated in Wikidata and associated with Wikidata Q IDs and a myriad of external identifiers including VIAF, ISNI, and regional identifiers. Adding workflows to manage entity authorities in BHL will improve collections metadata, bolstering the data model with reliable, legitimately unique identifiers.

With named entities established as a viable connection point between content and collections, another grant project, "Mining Biodiversity: Enriching Biodiversity Heritage with Text Mining and Social Media" focused on transforming BHL into a "next generation social-digital resource to facilitate the collaborative study and discussion of legacy biodiversity documents by a worldwide community."[29] Collaborators at the University of Manchester's National Centre for Text Mining, Missouri Botanical Garden center for Biodiversity Informatics, Dalhousie University's Big Data Analytics Institute and Ryerson University's Social Media Lab leveraged text mining, visualization methods, crowdsourcing, and social media to effectively serve its users with semantically enriched content to fully interlink and index the BHL corpus. This vision was to be accomplished through working packages comprised of OCR correction, corpus annotation, text mining, semantic searching, interactive visualizations, and social media initiatives. The most relevant sections for my work included OCR correction, semantic annotations, and text mining. As I experimented with ways to connect items across collections using their content and text, the annotation ontology identified by Mining Biodiversity was of particular interest. To ensure that the annotations are consistent specific ontologies and controlled vocabularies were selected and a schema was devised to save annotations as XML (XMI) files. Much of the literature was quite positive and supports that case for semantically enriching text in BHL: moving beyond scientific names and including context like habitat and geographic references, added people and corporate bodies, events, temporal expressions, and descriptions.

Text mining and entity extraction tools, however, depend on imperfect OCR that is generated automatically as part of BHL's digitization process. The Purposeful Gaming grant project was designed to correct OCR in order to improve the named entity recognition processes from Global Names, facilitate external text mining programs, and improve internal full-text search functions. In partnership with Dartmouth University's Tiltfactor Games Lab, BHL

[28] "IMLS Y3 Q3 calls," BHL Connecting Content Wiki (bhl.wikispaces.com/Connecting+Content). https://bhl.wikispaces.com/file/view/IMLS%20Y3%20Q3%20calls.pdf/443228562/IMLS%20Y3%20Q3%20calls.pdf

[29] Mining Biodiversity BHL Staff Wiki. https://biodivlib.wikispaces.com/Mining+Biodiversity

developed purposeful games to crowdsource OCR corrections. These games, Beanstalk and Smorball, were contrived of as fun, simple ways for humans to translate images of scanned words that computers cannot understand into indexable text that can be used for full-text searching.[30] Ultimately, the project did not work s intended and neither of the games were able to generate a critical mass of data corrections needed to develop an error matrix to automate further cleanup of the full BHL corpus. Perhaps different marketing could have improved responses, or the games could be reinstated until the needed number is reached, but when evaluated along with similar crowdsourcing projects in cultural heritage institutions it becomes clear that eliminating the connection between the game tasks and the actual BHL content is a real drawback. Improving resources and learning about relatively hidden materials are extremely compelling factors for many volunteers. Without providing direct access to the content, volunteers need a different type of motivation, and either the games themselves were insufficient or the premise that gaming can be motivating for this type of work is inherently flawed.

While OCR output is imperfect for published items in BHL, for handwritten documents it outputs mostly gibberish and is completely unusable. Ongoing crowdsourcing manuscript transcription programs including the Smithsonian Transcription Center, the Atlas of Living Australia's DigiVol platform, and collection based projects like the William Brewster diaries and correspondence at Harvard University's Museum of Comparative Zoology, the Joseph Grinnell field notes at the University of California Berkeley's Museum of Vertebrate Zoology, and the John Torrey Papers at the New York Botanical Garden and others invite volunteers to transcribe entire documents into machine readable text. Transcriptions have clear benefits including allowing users to quickly skim documents to determine content and usefulness, they reduce ambiguity, and can generate a consensus of what the content says. Crowdsourcing projects are also often cited as successful outreach tools. Volunteers are interested in contributing for a variety of reasons: often contributors need the information for their own research and many enjoy the creators' stories.

Furthermore, machine readable text transcriptions improve discovery tremendously when indexed and paired with a full-text search system, which allows users to find items according to their content in addition to metadata and catalog records or finding aid details. Indexed transcriptions can also be added to a faceted search and browse function that makes it easier to browse by commonly structured information such as taxonomic trees, identified people, institutions, and locations, document types, or even other publications. Transcriptions are also becoming vital for libraries and institutions that are increasingly adopting Collections as Data practices. For BHL, field notebooks contain a wealth of raw scientific data including unpublished observations, species occurrence records, habitat descriptions, weather reports with climate and meteorological data, phenological recordings and morphology information, sketches, and travel narratives. All of these potential entities, when researched across time and

[30] Purposeful Gaming BHL Staff Wiki.
https://bhl.wikispaces.com/Purposeful+gaming+IMLS+grant+project+page

across creators, often document local environmental conditions and may help identify gradual changes in biodiversity over many years.[31] Machine readable text transcriptions can turn raw and unevaluated primary sources into research data suitable for computational analysis that can be queried, sorted, and compared to each other and other data in external collections, data stores, knowledge bases, and information repositories.

In the realm of manuscript transcription crowdsourcing platforms there are essentially two types of projects: record based and document based.[32] Record based projects are more closely aligned with Citizen Science because they seek to extract tabular data from handwritten materials. 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.[33] Structured datasets generated from projects like Old Weather[34], FamilySearch Indexing,[35] [36] and the North American Bird Phenology Program and other specimen transcription citizen science projects[37] [38] follow this record based strategy in order to support research over very broad geographic and temporal scales.[39]

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.[40] 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

---

[31] DeVeer, Joseph M., Constance A. Rinaldo, and Linda Ford. "Primary Source Material in Science: The Importance of Archival Field Notes," Librarian e-Science Symposium 2013 http://escholarship.umassmed.edu/esciencesymposium/2013/posters/14.
[32] Brumfield, Ben. "What does it mean to 'support TEI' for manuscript transcription?" November, 2012. http://manuscripttranscription.blogspot.com/2012/11/what-does-it-mean-to-support-tei-for.html.
[33] *Ibid*.
[34] Blaser, Lucinda. 2014. Old Weather: Approaching Collections from a Different Angle. In *Crowdsourcing Our Cultural Heritage*, edited by Mia Ridge, 45–55. United Kingdom: Ashgate.
[35] Holley, Rose. "Crowdsourcing: How and Why Libraries Should Do It." *D-Lib Magazine* 16(3/4): 2010. doi: 10.1045/march2010-holley.
[36] Hanse, Derek, Jake Gehring, Patrick Schone, and Matthew Reid. "Improving Indexing Efficiency and Quality: Comparing A-B-Arbitrate and Peer Review." In Family History Technology Workshop. Brigham Young University. 2012
[37] Miller-Rushing, Abraham, Richard Primack, and Rick Bonney. "The History of Public Participation in Ecological Research." In *Citizen Science: Public Participation in Environmental Research*, edited by Janis L. Dickinson and Rick Bonney, 285–290. Ithaca: Cornell University Press, 2012. doi:10.1890/110278.
[38] Dickinson, Janis L., and Rick Bonney, eds. *Citizen Science: Public Participation in Environmental Research*. Ithaca: Cornell University Press, 2012.
[39] Mika, Katherine A., Joseph DeVeer and Constance Rinaldo, "Crowdsourcing Natural History Archives: Tools for Extracting Transcriptions and Data."
[40] Brumfield, Ben. "What does it mean to 'support TEI' for manuscript transcription?"

conventions and may include toolbars to encourage consistency.[41] 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 data for 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.[42]

Early manuscript transcription crowdsourcing programs were largely collections based and were tied to specific manuscript collections of significant research value or interest or that are at high risk for long term physical preservation. These projects depended on high levels of archivist curation and were designed for smaller programs for which significant manual data and community management practices were plausible. Workflows for these projects often require significant time investments from collections staff, and do not typically scale well. Even with the expertise of a dedicated community or program manager it is unsustainable for very large collections. Fortunately, as this domain of manuscript transcription continues to develop and since crowdsourcing has proved to be a reliable method for generating machine readable text, platforms are evolving to meet the demands of larger digital collections by incorporating citizen science practices and adding markup, tagging, and semantic annotation features to give this information structured meaning and connect it to larger knowledge bases. Fortunately, there are several high quality platforms that can generate text data in support of an interoperable linked data graph.

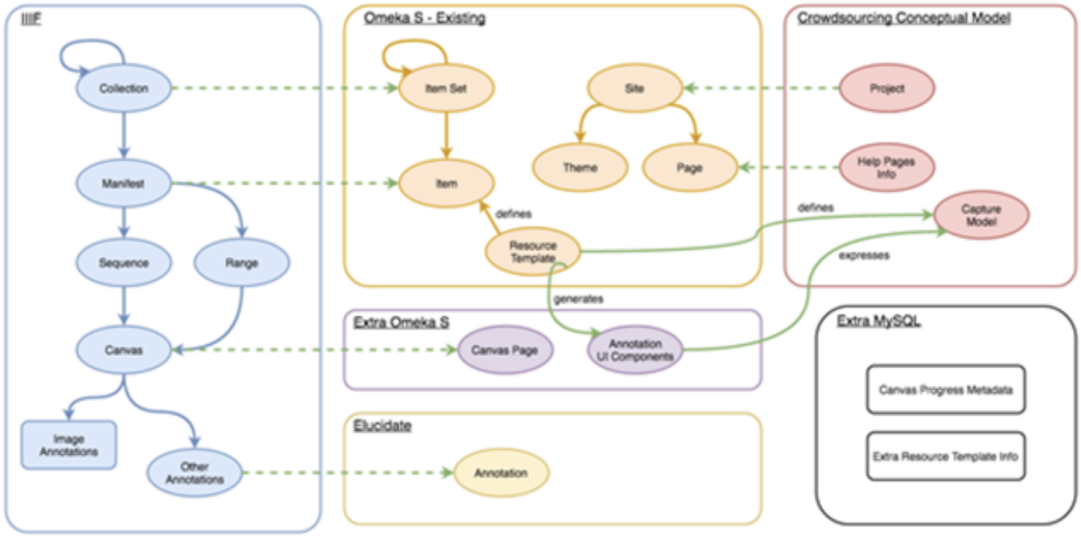
The Smithsonian Institution's Transcription Center, Brumfield Labs' FromThePage software, The Atlas of Living Australia's DigiVol platform, and The Zooniverse are all terrific options for partner institutions to build transcription projects. A paper from earlier in the NDSR residency, "Crowdsourcing Natural History Archives" offers a comparison of these tools that identifies specific strengths and weaknesses of each. However, an institution's project goals, scope of collections, item types, and resources available for engaging in transcription projects will be important to consider in a selection process in addition to general platform information and evaluation of successful, archetype projects. One significant barrier to relying on crowdsourcing for large scale projects is scalability. Reported statistics for BHL and SI Field Book and Field Note projects include 3,062 digitized items totaling 377,965 pages,[43] which represents the holdings of only a handful of BHL partner institutions. Transforming images of handwritten documents into meaningful machine readable data requires many volunteer hours of content creation and data entry and many staff hours of pre- and post-processing and project management. Transitioning to a corpus-wide transcription system with continued support for partner contributions would allow BHL to leverage its distributed staff model within a single platform.

Accordant with other Residents' recommendations to transition from the Internet Archive to a BHL owned image delivery system that remains compatible with the International

[41] *Ibid.*
[42] Mika, Katherine A., Joseph DeVeer and Constance Rinaldo, "Crowdsourcing Natural History Archives: Tools for Extracting Transcriptions and Data."
[43] BHL Staff Call. https://bhl.wikispaces.com/bhlstaffcalldec2017

Image Interoperability Framework (IIIF), my recommendation is to consider an image viewer that supports crowdsourced annotations which can be used to transcribe handwriting, correct OCR, and semantically enrich text data with linked data ontologies and controlled vocabularies. The National Library of Wales is partnering with Digerati digital consultancy firm to develop an open source crowdsourcing module that combines IIIF APIs and the W3C Web Annotation Model to produce an extensible platform that will be published in early 2018, and is currently available in collaboration with the Digital Library Cloud Services[44] offered by Digerati.[45] [46] The diagram below [Fig. 5] illustrates how the IIIF image server, BHL book viewer/IIIF image viewer (OmekaS in the diagram), annotation server (Elucidate), and BHL database might all interact with an additional Crowdsourcing Conceptual Model layer to generate.

Figure 5: Crowdsourcing Annotations Model[47]

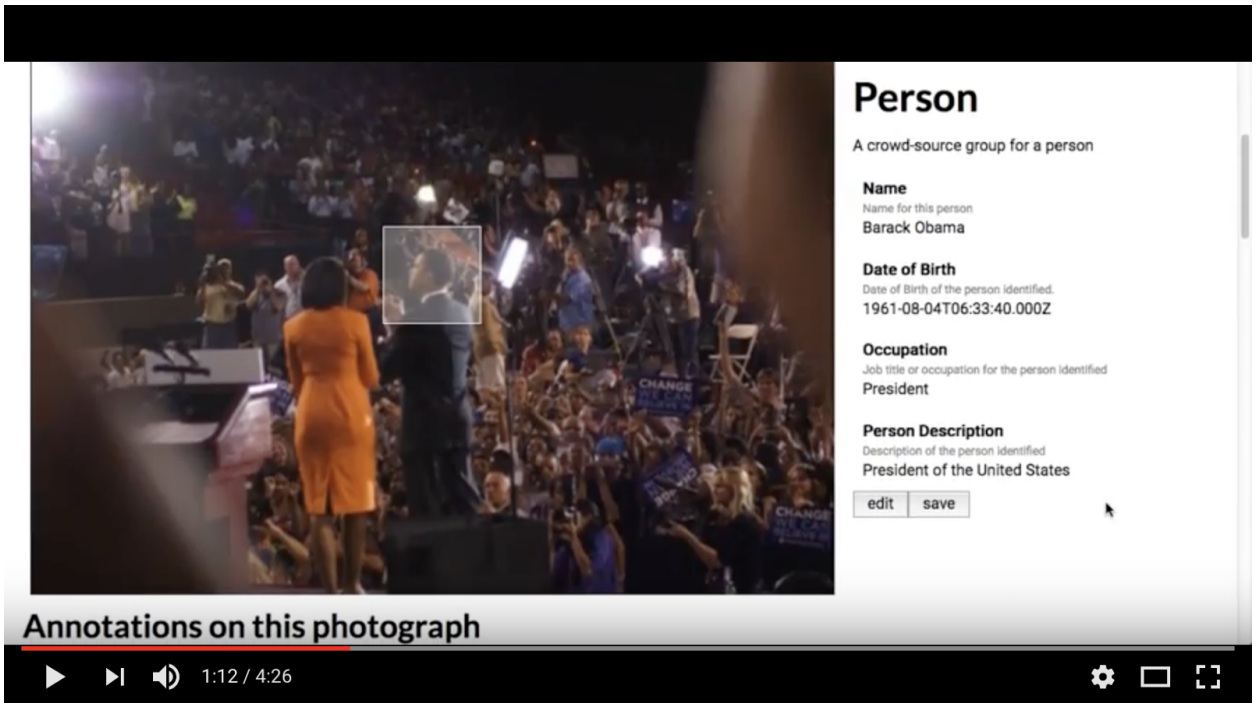


The technical infrastructure in this model stores page images in an image server which communicates with the image viewer via API to write the image onto a canvas each time the page is loaded with varying requirements (deep zooms, color or grayscale enhancements, etc). An annotation server stores data as W3C web annotations[48] and similarly communicates with the image viewer to write annotations, user comments, transcriptions, or OCR text onto the page. The crowdsourcing module from Digerati's DLCS applies a conceptual model to a IIIF

[44] https://dlcs.info/index.html
[45] Robson, Glen. "IIIF Crowd Sourcing Platform." Slides from presentation: National Library of Wales IIF Web Annotation Crowdsourcing at IIIF Conference 2017, Vatican City, Italy. https://lonewolflibrarian.wordpress.com/2017/06/12/national-library-of-wales-iiif-web-annotation-crowdsourcing/ .
[46] James, Rhian. "A Crowdsourcing Platform for Wales." Nationals Library of Wales blog. Februrary 15, 2017. https://www.llgc.org.uk/blog/?p=14496.
[47] "A Crowdsourcing Platform for Wales." Digerati. 2017. https://digirati.com/updates/insights/a-crowdsourcing-platform-for-wales/
[48] W3C Web Annotations are replacing Open Annotations as the annotation standard for the web. See https://www.w3.org/TR/annotation-model/ for more information.

compatible repository (OmekaS in the case of the National Library of Wales) to design a project and define volunteer tasks that are contributed in the form of annotations [Fig. 6]. The platform can then execute the model by generating navigation, rendering pages, and rendering the UI components for model capturing.[49] These model components support the use of controlled vocabularies and biodiversity ontologies that could define relationships between taxa and descriptions, locations, added dates and people, etc.[50] Furthermore, the module requires merely the IIIF image URL to add annotations and can pull URLs from any IIIF compatible resource. Digerati has indicated that it would be feasible to develop a crowdsourcing program for BHL using the current Internet Archive image delivery system.

Figure 6: National Library of Wales Crowdsourcing Demo[51]



The existing technical framework and infrastructure for this system can be implemented independently of BHL's database structure and catalog, but is still able to pull specific item, title, and page metadata into the image viewer. BHL can continue to source MARC records from contributing libraries for the catalog, while the book viewer pulls scientific names from the GNRD index, Encyclopedia of Life (EOL) links from the EOL API, and annotations from the annotation server which can include OCR djvu files, transcriptions, and user comments and annotations. Some development will be required to either edit the existing book viewer ito recognize a different IIIF image delivery system that supports annotations, or to add features to a new open source annotation-compatible image viewer, such as Mirador,

[49] "A Crowdsourcing Platform for Wales." Digerati. 2017. https://digirati.com/updates/insights/a-crowdsourcing-platform-for-wales/
[50] "International Image Interoperability Framework for Natural History Collections." Digerati webinar organized for the Royal Edinburgh Botanic Garden and CETAF. https://youtu.be/8Op1-vw-0Vw.
[51] Full video available at: https://youtu.be/oSCdQgf9nSI.

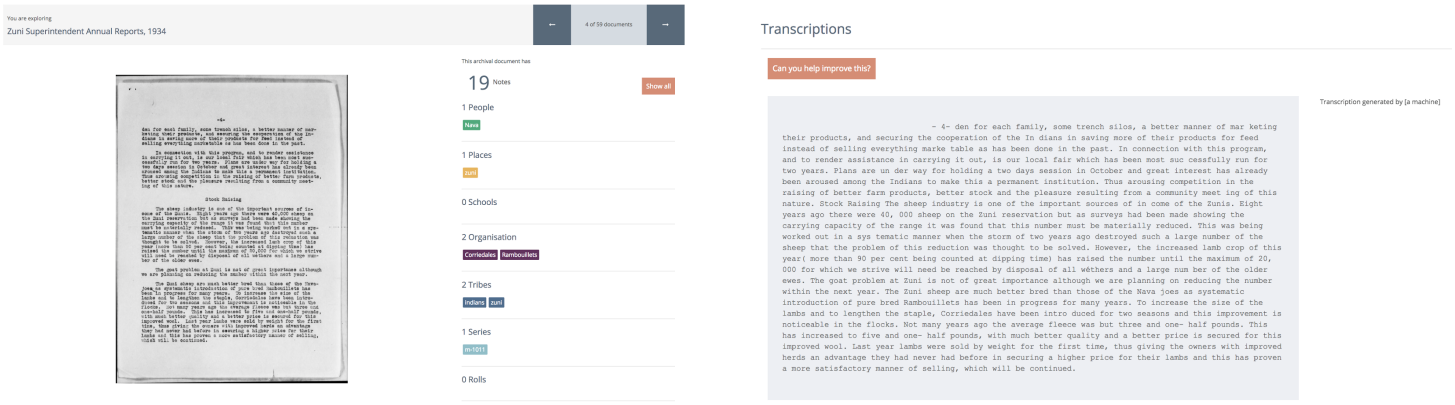
that can display links to scientific name bibliographies and EOL links. Furthermore, as the biodiversity knowledge graph continues to expand and adopt linked data technologies to connect library collections, museum specimen collections, taxa, nomenclatures, authors, creators, and people, organizations and corporate bodies including funders, locations, traits, genomic and phylogenetic data, et al., BHL can add graph nodes for literature and special collections by publishing IIIF annotations as JSON-LD. Future development towards a graph database and linked open data model for BHL collections metadata would be compatible with the W3C web annotations on item images generated from the crowdsourcing module. Adopting a linked data strategy for opening crowdsourced annotations to the Semantic Web does not conflict with the current BHL data model altogether, but will likely require some updates and changes to the system to support a new module.

A portal-based platform must also support integrating transcriptions and their encoded data from partner institutions. Fortunately, converting and transforming data outputs from the most widely used platforms including FromThePage, the Zooniverse, DigiVol, and the Smithsonian Transcription Center should be somewhat straightforward. FromThePage is developing documentary editing for IIIF which supports IIIF standards in FromThePage and enables collaborative transcription and annotation of material on any IIIF enabled repository.[52] While this largely refers to the interaction between FromThePage and an institution's IIIF image server, the structured TEI (Text Encoding Initiative) XML output combined with the image's pixel location information will facilitate its expression in a W3C web annotation layer in the BHL book viewer. Similarly, the Smithsonian Transcription Center's XML output can also be written into an annotation layer provided that the pixel image target data is preserved. The Zooniverse exports data from its Project Builder module as JSON (within CSV files) and can be parsed accordingly. Bespoke Zooniverse collaborations may improve interoperability with IIIF and W3C Web Annotation models.

Incorporating OCR files as annotation layers in an image viewer reveals further crowdsourcing opportunities for correcting texts [Fig. 7]. Selected titles and items can be opened up for volunteers to correct and improve text data. While the Purposeful Gaming project failed to generate a sufficient amount of data, its plan to develop an error matrix from which corrections can be applied using regular expressions to the BHL corpus is still achievable. Creating several OCR corrections crowdsourcing projects similar to successful manuscript transcription projects and leveraging BHL's robust outreach program to promote them could be a feasible strategy for completing the error matrix.

[52] http://www.brumfieldlabs.com/

Figure 7: Indigenous Digital Archive crowdsourcing portal[53]



As crowdsourcing is brought into the BHL portal, validation mechanisms will be required to vet data the volunteers contribute. While there is some concern for vandalism in these types of projects, there is little evidence that it has become a pervasive or widespread problem in GLAM crowdsourcing projects. Indeed, the professional and ethical obligation to open collections to increasingly diverse communities of users should outweigh staff apprehensions over potential digital iniquities. However, BHL does have a valid interest in mediating taxonomic disputes and disagreements between users and identifying mistakes as part of its crowdsourcing program. The crowdsourcing module includes a feature in the W3C annotation life cycle that embargos user generated content in an administrative dashboard for staff or expert moderation. There are also after-the-fact options that allow administrators to delete or hide annotations in lieu of or in addition to approving annotations.

Finally, the DLCS IIIF crowdsourcing module is an optimal choice for the organizational structure of BHL. While adding a staff member dedicated to managing transcription projects and community engagement would likely improve the quality of annotations, a committee of staff members from several institutions could reasonably expect to maintain and administer the platform collaboratively. Plugging a new feature into BHL's successful distributed staff model increases the likelihood of institutional buy-in. In an organizational environment where several staff members from global BHL partners are extremely interested in a transcription solution, the support for creating this type of committee could be quite high. Applying this committee model to a transcription platform also increases the scalability of crowdsourcing transcriptions. An interested and committed group of skilled librarians and archivists are more likely to successfully manage increased and variable levels of volunteer interactions, promote particular collections and field books, possess expert knowledge of collections and creators, and handle content validations and authorizations better than a single person at a single institution. Additionally, adding a significant number of staff hours and institutions to the

[53] "Zuni Superintendent Annual Reports, 1934." Indigenous Digital Archive. https://omeka.dlcs-ida.org/s/ida/iiif/manifest/1517/5a407545e7900

program communicates the high level of investment by BHL. Users and volunteers may be more interested in participating when BHL as a whole appears seriously committed to the program's success.

Several organizations have implemented Digerati's DLCS program and its crowdsourcing module including the Royal Society, the National Library of Wales, and the Indigenous Digital Archive. The Royal Society digitized and uploaded the archive of *The Philosophical Transactions of the Royal Society*, the oldest continuously published science journal in the world, asa pilot project to transcribe, tag, and comment on the material and provide new ways of linking items together.[54] For this project, the DLCS ensured that the technical approach allowed the digitized source material and the user generated content (transcriptions, tags, comments, and bookmarks) were made available via open standards for later reuse - IIIF for the archive's material, and the W3C Web Annotation Data Model and Open Annotations for the user supplied content. The National Library of Wales collaboration aimed to build a reusable, open source platform to carry out various crowdsourcing projects on its rich collections. The platform is designed to create crowdsourcing projects using IIIF library resources and add user generated annotations that are compliant with open standards.[55] The Indigenous Digital Archive at the Museum of Arts and Culture in Santa Fe builds on the crowdsourcing module to include OCR corrections, named entity extraction, and other text processing elements.[56] Corrections and named entities are identified by volunteers, staff, and machines according to controlled vocabularies, which are refined and managed in a Topic Manager tool. The automated and human-assisted processes rapidly build a navigable corpus of documents that uses the recognized and refined entities to encourage exploration and discovery.[57] Each of these programs successfully demonstrates the extensibility and flexibility of the DLCS IIIF Crowdsourcing Module and its power to enhance discovery, access, data collection, and reuse of library items and their content and metadata.

Further investigation will obviously be required in order to determine feasibility, however, semantic enrichment of collections content data and metadata is being rapidly adopted as Best Practices for digital library and archives collections. Tagging features were added to transcription platforms as crowdsourcing projects were implemented. Navigating items according to their content became useful as machines became able to read digitized content. As institutions added increasing amounts of text data to their repositories, encoding became a natural extension to help manage a large amount of digital information. Establishing schemas and vocabularies to help machines understand this data brought features supporting computational research by users across larger and larger datasets. IIIF and the W3C Web Annotation Data Model have identified useful standards for making digital collections data

---

[54] "The Royal Society Journal Collection: Science in the Making." DLCS Case Study. 2017. https://dlcs.info/case-study-rs.html.
[55] "A crowdsourcing platform for Wales." DLCS Case Study. 2017. https://dlcs.info/case-study-nlw.html.
[56] Naruta, Anna. "Technology Behind the IDA Toolkit -- A roundup of links and videos." Indigenous Digital Archive Blog. October 10, 2017. http://blog.indigenousdigitalarchive.org/post/2017/10/10/Technology-behind-the-IDA-toolkit-a-roundup-of-links-and-videos.
[57] "A crowdsourcing platform developed for the Museum of Indian Arts and Culture in Santa Fe." DLCS Case Study. 2017. https://dlcs.info/case-study-ida.html.

FAIR.[58] Combining these interoperable image and data standards with Linked Open Data technologies will allow BHL to better connect its users to its own content as well as data from other nodes on the emerging biodiversity knowledge graph.

As BHL investigates methods for adding to and restructuring its data and metadata architecture, it should consider trends in library technology in addition to methods for supporting various types of computational research for its text collections. It would be a missed opportunity and likely create future problems to develop a long term program for transcribing documents and correcting OCR output in ways that cannot or will not be used for the type of research that scientists and researchers are increasingly engaged in. BHL can take advantage of the work that other transcription projects have already done in learning about about how to support more traditional "close reading" analyses of manuscript material. However, as a digital library focused on scientific inquiry, BHL is well situated to become a leader in the Collections as Data movement by developing programs and an updated data model to support computational analysis and distant reading of texts.

[58] Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Comments and Opinion. Scientific Data, March 15, 2016. https://doi.org/10.1038/sdata.2016.18.

Bibliography

"A crowdsourcing platform developed for the Museum of Indian Arts and Culture in Santa Fe." DLCS Case Study. 2017. https://dlcs.info/case-study-ida.html.

"A Crowdsourcing Platform for Wales." Digerati. 2017. https://digirati.com/updates/insights/a-crowdsourcing-platform-for-wales/.

Berry, David M. "The Computational Turn: Thinking About the Digital Humanities." *Culture Machine* 12, February 18, 2011. http://www.culturemachine.net/index.php/cm/article/view/440.

Bik, Holly M. "Let's rise up to unite taxonomy and technology." *PLOS Biology*. 15(8): e2002231. August 18, 2017. https://doi.org/10.1371/journal.pbio.2002231.

Blaser, Lucinda. 2014. Old Weather: Approaching Collections from a Different Angle. In *Crowdsourcing Our Cultural Heritage*, edited by Mia Ridge, 45–55. United Kingdom: Ashgate.

Brumfield, Ben. "What does it mean to 'support TEI' for manuscript transcription?" November, 2012. http://manuscripttranscription.blogspot.com/2012/11/what-does-it-mean-to-support-tei-for.html.

Computational Archival Science Workshop, IEEE Big Data 2017. http://dcicblog.umd.edu/cas/ieee_big_data_2017_cas-workshop/.

"Construction of biodiversity knowledge graphs driven by federated text mining tools." *Biodiversity Information Standards, TDWG 2017*. Ottawa, Ontario, October 2017. https://biss.pensoft.net/browse_user_collection_documents?collection_id=44.

DeVeer, Joseph M., Constance A. Rinaldo, and Linda Ford. "Primary Source Material in Science: The Importance of Archival Field Notes," Librarian e-Science Symposium 2013 http://escholarship.umassmed.edu/esciencesymposium/2013/posters/14.

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

Faulwetter, Sarah, Evangelos Pafilis, Lucia Fanini, Nicolas Bailly, Donat Agosti, Christos Arvanitidis, Laura Boicenco, et al. "EMODnet Workshop on Mechanisms and Guidelines to Mobilise Historical Data into Biogeographic Databases." *Research Ideas and Outcomes* 2 (June 7, 2016): e9774. https://doi.org/10.3897/rio.2.e9774.

Fischer, Richard T. "Connecting Content: A Collaboration to Link Field Notes to Specimens and Published Literature." Society of American Archivists. May 2012. https://www2.archivists.org/sites/all/files/RichardFischer-SAA-ResearchPaper-2011.pdf

"Foundations to Actions: Extending Innovations in Digital Libraries in Partnership with NDSR Learners." Institute for Museum and Library Services, Laura Bush 21st Century Librarian Program. Award number: RE-40-16-0082-16. https://www.imls.gov/grants/awarded/RE-40-16-0082-16.

Hanse, Derek, Jake Gehring, Patrick Schone, and Matthew Reid. "Improving Indexing Efficiency and Quality: Comparing A-B-Arbitrate and Peer Review." In Family History Technology Workshop. Brigham Young University. 2012.

Holley, Rose. "Crowdsourcing: How and Why Libraries Should Do It." *D-Lib Magazine* 16(3/4): 2010. doi: 10.1045/march2010-holley.

"IMLS Y3 Q3 calls," BHL Connecting Content Wiki (bhl.wikispaces.com/Connecting+Content). https://bhl.wikispaces.com/file/view/IMLS%20Y3%20Q3%20calls.pdf/443228562/IMLS%20Y3%20Q3%20calls.pdf.

"International Image Interoperability Framework for Natural History Collections." Digerati webinar organized for the Royal Edinburgh Botanic Garden and CETAF. https://youtu.be/8Op1-vw-0Vw.

James, Rhian. "A Crowdsourcing Platform for Wales." Nationals Library of Wales blog. February 15, 2017. https://www.llgc.org.uk/blog/?p=14496.

Marciano, Richard, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. "Archival Records and Training in the Age of Big Data." In *Advances in Librarianship - Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education*. Eds.: Lindsay C. Sarin, Johnna Percell, Paul T. Jaeger, & John Carlo Bertot. In Press. http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/Marciano_Kurtz_et-al-Archival-Records-and-Training-in-the-Age-of-Big-Data-final-1.pdf

Miller-Rushing, Abraham, Richard Primack, and Rick Bonney. "The History of Public Participation in Ecological Research." In *Citizen Science: Public Participation in Environmental Research*, edited by Janis L. Dickinson and Rick Bonney, 285–290. Ithaca: Cornell University Press, 2012. doi:10.1890/110278.

Mika, Katherine A., Joseph De Veer, and Constance Rinaldo. "Crowdsourcing Natural History Archives: Tools for Extracting Transcriptions and Data." *Biodiversity Informatics* 12, no. 0 (November 14, 2017). https://doi.org/10.17161/bi.v12i0.6646.

Naruta, Anna. "Technology Behind the IDA Toolkit -- A roundup of links and videos." Indigenous Digital Archive Blog. October 10, 2017. http://blog.indigenousdigitalarchive.org/post/2017/10/10/Technology-behind-the-IDA-toolkit-a-roundup-of-links-and-videos.

Padilla, Thomas, Laurie Allen, Stewart Varner, Patricia Hswe, Elizabeth Roke, and Sarah Potvin. "Always Already Computational: Library Collections as Data." Institute for Museum and Library Services. National Leadership Grants for Libraries, National Digital Platform Project Proposal. 2016.

Page, Roderic D. M. "BioNames: Linking Taxonomy, Texts, and Trees." *PeerJ* 1 (October 29, 2013): e190. https://doi.org/10.7717/peerj.190.

Page, Roderic D. M. "Towards a Biodiversity Knowledge Graph." Conference Keynote Presentation, Biodiversity Information Standards (TDWG), October 2, 2017. http://idigbio.adobeconnect.com/p2ul6gfd20os/

Robson, Glen. "IIIF Crowd Sourcing Platform." Slides from presentation: National Library of Wales IIF Web Annotation Crowdsourcing at IIIF Conference 2017, Vatican City, Italy. https://lonewolflibrarian.wordpress.com/2017/06/12/national-library-of-wales-iiif-web-annotation-crowdsourcing/.

"Semantic Web." World Wide Web Consortium. https://www.w3.org/standards/semanticweb/

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." Comments and Opinion. Scientific Data, March 15, 2016. https://doi.org/10.1038/sdata.2016.18.

"Why Linked Data?" Linked Data for Libraries. https://www.ld4l.org/linked-data.