Final Report

Foundations to Actions: Extending Innovations to Digital Libraries in Partnership with NDSR Learner

Content Analysis

Alicia Esquivel, Chicago Botanic Garden

November 2017

Executive summary **3**

Introduction **4**

Review of Content Analysis **5**
Previous content analyses at BHL 5
Selected Literature review 5

Case Studies **6**
Temporal Analysis 6
Taxonomic Analysis 7
Topical Analysis 9
Geographic Analysis 12
Statistical Analysis 15

Recommendations **18**

Role at Lenhardt Library **18**

Conferences **19**

Appendix **22**
Bibliography 22
Technical documents 23
Temporal Technical Document 23
Taxonomic Technical Document 23
Geographic Technical Document 26
Statistical Analysis Technical Document 26
List of Posters and Presentations 27
Links to blog posts 28
From Resident blog, authored by me: 28
From Resident blog, collaboratively authored by all Residents: 28
From BHL blog, collaboratively authored by all Residents: 29
From CBG Library blog, authored by me: 29

Executive summary

This report is an overview of one of the National Digital Stewardship Residency[1] (NDSR) projects from the Foundations to Actions: Extending Innovations to Digital Libraries in Partnership with NDSR Learners cohort.[2] In addition to project work, I will describe the residency experience in full including functions and responsibilities served as a resident and as an employee of the host institution.

From January 2017 to November 2017, I worked on a content analysis project for the Biodiversity Heritage Library (BHL) while hosted at the Chicago Botanic Garden's Lenhardt Library.[3] The two main questions I sought to answer in my content analysis were: what has been digitized by BHL? And what has not yet been digitized by BHL? The answers to these questions can help focus future digitization efforts by BHL and inform digital libraries on how to perform a content analysis of a digital collection.

Using bibliographic metadata and OCR'd text, I analyzed BHL's collection temporally, geographically, topically, and taxonomically. In order to calculate the amount of biodiversity literature not yet digitized by BHL, I performed statistical analysis. These case studies can be scaled up and replicated for BHL and other digital libraries to assess content coverage of a collection. Technical documentation, including links to code, links to raw data, and links to results can be found in the appendix.

The residents created a Wordpress blog and published updates throughout the year to keep the public informed on our projects. We also wrote guest blog pieces for the BHL blog every other month and some residents, myself included, wrote blog pieces for host institutions' blog. Links to these posts can be found in the appendix.

Additionally, I attended regional, national, and international conferences as a representative of BHL and the NDSR program. Links to presentations and posters from these conferences can be found in the appendix.

---

[1] https://ndsr-program.org/

[2] http://biodivlib.wikispaces.com/Extending+Innovations+in+Digital+Libraries+in+Partnership+with+NDSR+Learners

[3] https://www.chicagobotanic.org/library

Introduction

The National Digital Stewardship Residency program (NDSR) places new graduates of library and information science in projects to foster the development of digital stewardship professionals. Each resident is paired with a mentor at a host institution to guide the project and professional development of the resident. The Foundations to Actions: Extending Innovations to Digital Libraries in Partnership with NDSR Learners project placed 5 residents in geographically dispersed host institutions to develop best practices and recommendations for a next generation digital library.

As the resident hosted at Chicago Botanic Garden, with mentor Leora Siegel, Director of the Lenhardt Library, I have been responsible for conducting content and gap analysis of the Biodiversity Heritage Library (BHL). The goals of this project were to analyze the "quantity of literature underpinning the field of biodiversity, the amount of that literature in the public domain, the representation of each discipline (delineated by taxon group) within BHL, an exploration of methodologies to scope the collections, and areas where BHL may target development to better serve the research population."[4]

Throughout the year I worked with fellow residents to complete my project. We communicated regularly through Slack, a chat interface, and held video meetings every week on Google Hangouts. We each gave updates to our mentors on a regular basis and met as a full resident and mentor group twice a month via conference call. The residents were also embedded in the larger BHL community and attended conference calls including staff meetings, Collections Committee meetings, and Technical Team meetings. During the first month of the residency, all mentors and residents attended a "Boot Camp" program held at the Smithsonian's National Museum of Natural History where we were introduced to the organization of BHL, its collection development policies, and its workflow processes for uploading material through Internet Archive.

I started my research by reviewing previous content analyses performed at BHL by the Collections Committee. Then I conducted a literature review about content analyses in general, content analyses for digital libraries, and content analyses for a biodiversity collection. Based on this research, I decided to analyze the BHL corpus across four categories best fit for biodiversity literature: time, taxonomy, topic, and geography. I then explored options for analyzing the collection by these categories using the available data which includes bibliographic metadata and OCR'd full text. As a result, I have four case studies that can be expanded or improved upon for further analysis of the BHL corpus and a fifth case study that uses statistical analysis to quantify the amount of literature underpinning the field of biodiversity.[5]

---

[4] https://www.imls.gov/sites/default/files/re-40-16-0082-16_proposal_documents.pdf

[5] All metadata was collected in March 2017 and reflects the holdings of the collection at that time.

Review of Content Analysis

Previous content analyses at BHL

Content analysis has been a priority of BHL's Collection Committee since its founding. The requirements for a content analysis according to the committee's wiki are "To determine the extent of the biodiversity literature universe and how much BHL has fulfilled / left to scan?" And "What are areas of BHL corpus strength and weakness?: [in regard to] bibliographic subject matter, taxonomic subject matter, geographic region and language."[6]

After BHL's founding in 2006, members of the Collection Committee began looking at ways to prioritize scanning and reduce duplications from the 10 member organizations. The OCLC Collection Analysis tool was purchased and all member institutions' collection records were imported.[7] This enabled members to move forward with scanning efforts by assigning appropriate subjects to member institutions.

In 2010, the Collections Committee estimated the size of core biodiversity literature in order to gauge BHL's progress in scanning and making available biodiversity literature. Two large bibliographies, *Taxonomic Literature, 2nd ed. (TL2)* and *Botanico-Periodicum-Huntianum* were used to estimate the number of monographs and serials published on botany and the page count. This page count was compared to the estimate of botanical species. This was then extrapolated to animals and fungi.

Most recently, in 2015 and 2016, Jackie Chapman and Robin Everly of Smithsonian Libraries created a bibliography of ferns from *Fern Books in English Published before 1900*, by keying in bibliographic information by hand. They searched for all books, not serials, from this bibliography in BHL to gauge coverage of fern literature in BHL. They also created a controlled vocabulary of fern species from Catalogue of Life and compared it to scientific names in BHL.

Selected Literature review

In addition to BHL's previous work in content analysis, I researched methods pursued by other digital libraries and biodiversity collections. In September 2016, the Global Biodiversity Information Facility (GBIF) published a best practices guide for conducting a content gap analysis. While the guide focuses on biodiversity data, not literature, the framework has proven to be useful for this BHL literature analysis. The basic basic steps for conducting a data gap analysis are:

"1. Scoping the analysis and setting the expectations
2. Assessing the universe of accessible data
3. Identification of data gaps
4. Synthesis and dissemination of the outcomes

---

[6] From "Final document on recording questions we want to ask of the BHL collection for a Collections Analysis project," BHL private wiki. https://docs.google.com/document/d/1GG9vbRIV5bWKTHpfHKAJKxW4l91JfkUc8_-nlgeber8/edit

[7] https://bhl.wikispaces.com/OCLCAnalysisconcerns

5. Prioritization of gap-closing, demand-driven data discovery and publishing activities
6. Evaluation of the DGA exercise."[8]

I executed step 1 by researching previous analyses and unpacking project goals. Step 2 required me to consider the types of data available for content analyses which includes bibliographic metadata and OCR'd full text files. I used the information from steps 1 and 2 to work towards step three, identifying gaps, by conducting 4 case studies that analyzed the collection of BHL temporally, taxonomically, topically and geographically. Each of these case studies can be replicated by BHL for further analyses.

Case Studies

Temporal Analysis

In order to see where the collection strengths and weaknesses were over time, I graphed each item from BHL by year. I used metadata from the BHL data export page, specifically the item and title tables.[9] Because of the different cataloging practices at contributing institutions, not all items have a date available in the date field which is why I also needed to look at the title table.

Using Python, I merged the title and item tables, giving each item three potential date fields including: year, startyear, and endyear.[10] If the year field contained a valid date, that was used as the item's date, if year was blank, the Python script looks at startyear and endyear and represents the date as a range. The range is weighted it as a percentage of the item (so if an item had a start date of 1880 and end date of 1885, there was a range for those five years, and each year got 1/5 of an item represented on the chart). The chart was created using matplotlib, a Python library for 2D plotting. [11]

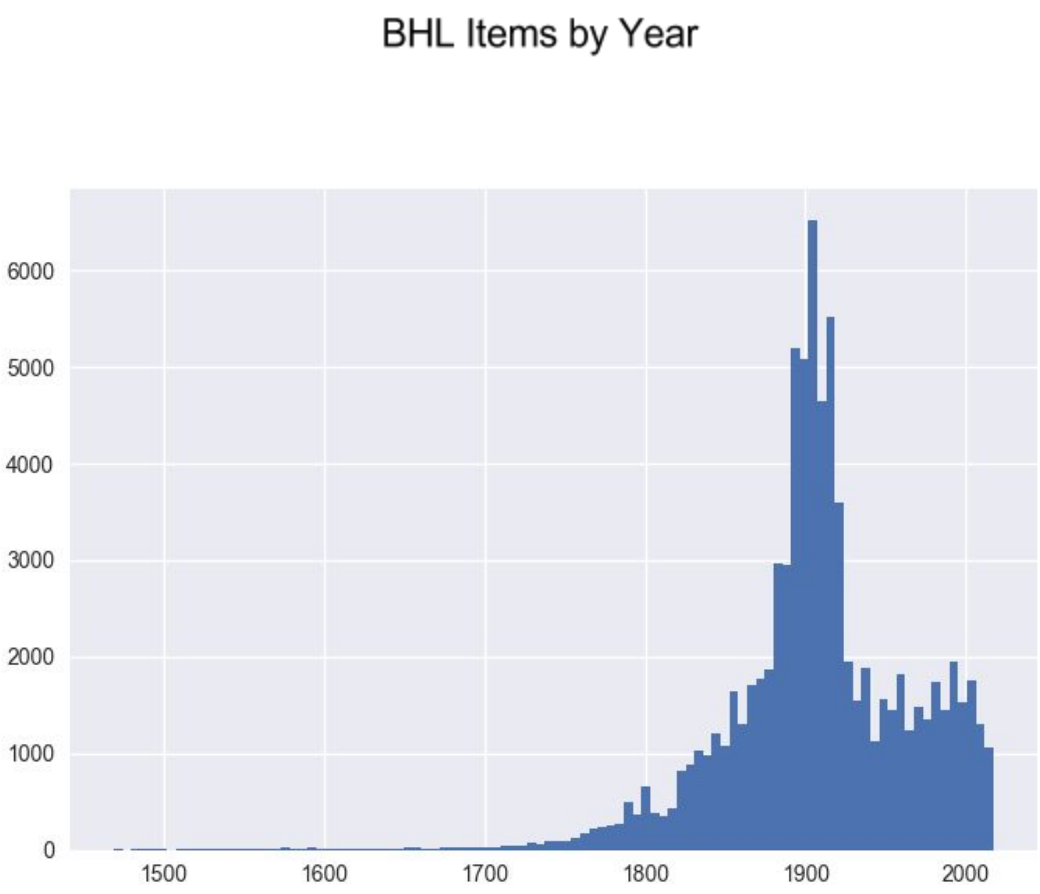
---

[8] Arturo Ariño, et al. *Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders.* (GBIF, 2016).

[9] http://biodivlib.wikispaces.com/Data+Exports

[10] Python script available here https://osf.io/3yuhr/; merged tables available here https://osf.io/2bx6p/

[11] https://matplotlib.org/



These results show a wide distribution of texts with a peak around the turn of the 19th century and a fall off around the 1920s due to copyright and public domain. This graph can be used as a snapshot of the collection and should be compared to future analyses to mark growth. Additionally, this graph can be used to make comparisons to other collections to look for similar or differing trends in publications.

Taxonomic Analysis

One of the most valuable access points into a biodiversity collection is through scientific names.[12] When scientific names are indexed, users such as scientists and taxonomists, BHL's largest user group, can quickly locate relevant content pertaining to specific taxonomic ranks for research. BHL uses Global Names Recognition and Discovery (GNRD) services to identify and extract scientific names from the OCR'd text.[13] Advanced searches can be made in BHL using scientific names and users are directed to pages in the corpus where the names are found. Additionally, scientific names in BHL link out to Encyclopedia of Life pages. While scientific

[12] Dimitry Mozzherin, et al. *Challenges with using names to link digital biodiversity information.* (Biodiversity Data Journal, 2016).
[13] http://resolver.globalnames.org/

names in BHL are indexed, they have no internal structure or organization. In order to gain understanding about content coverage of BHL based on taxonomy, I matched the list of BHL's scientific names to a taxonomic backbone and filtered names into their taxonomic kingdom.

The list of every occurrence of scientific name in BHL and it's corresponding pageID number is available for download as a TXT file from the BHL data export page as pagename.txt. This huge file (6,321,666 KB) is what I used to analyze the BHL collection.

After downloading pagename.txt, I created a dictionary in Python of every unique name from this list with a count of every time it occurred.[14] This resulted in 3,732,986 unique scientific names in BHL.[15] However, this number is inflated due to a lack of standardization in scientific name formats in the BHL corpus. For example, GNRD recognized and extracted: "Echinaster serpentarius," "Echinaster serpentarius Müller & Troschel," and "Echinaster serpentarius Müller and Troschel" from the BHL corpus and the Python code for creating a dictionary is representing these as 3 unique scientific names because of differences in these strings when they actually represent the same name. Furthermore, GNRD analysis is dependent upon uncorrected OCR results with varying quality.

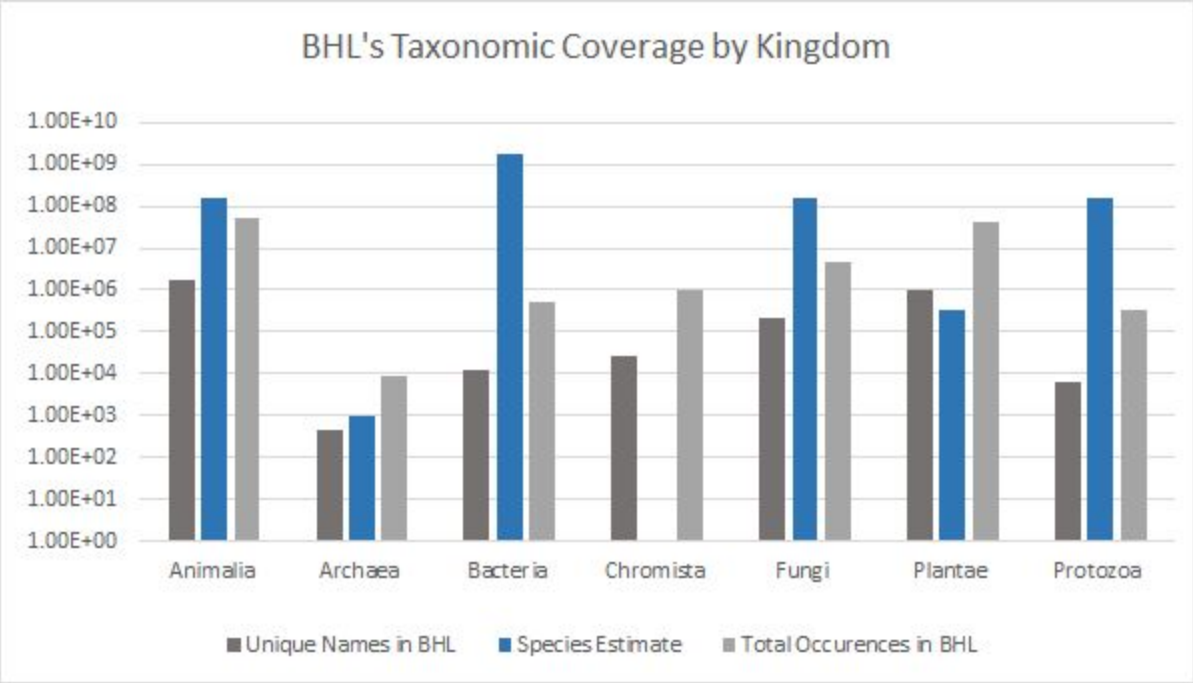
After creating a list of unique names, I needed to organize or filter them in an intelligible way. GNRD allows you to choose a data source, or taxonomic backbone, to run your names against and generates a "classification path" for a scientific name which places a name on the tree of life. I tested multiple sources to see which gave the best results and chose to compare BHL names to Catalogue of Life.[16] With the help from GNRD developer, Dima Mozzherin, I was able to run lists of names in csv format, 100,000 names at a time. I wrote a script in Python to break up the dictionary with 3.7 million names into smaller lists of 100,000 names each and manually uploaded each list to GNRD resolver.[17] Each list of 100,000 names took approximately one hour and 45 minutes to process. Full results from these 38 files are available as CSV files.[18]

To analyze the results I first dropped any names that did not match or resolve to the Catalogue of Life (18%) and removed duplicates that occurred when input names matched to multiple source names. I then extracted kingdoms from the classification paths and counted the unique names in each kingdom and multiplied each kingdom by the counts. This was all done with a script written in Python.[19]

I plotted these results against each other and against a species estimate.[20] (Note, the reference used for species estimate does not include an estimate for Chromista kingdom).

---

[14] Python script available here https://osf.io/s7fdw/
[15] Dictionary created found here https://osf.io/9nucr/
[16] https://osf.io/fue9t/
[17] Python script available here https://osf.io/g73af/
[18] https://osf.io/tfhg3/
[19] https://osf.io/xnu8b/
[20] Brendan B. Larsen, et al., *Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life.* (The Quarterly Review of Biology, 2017).



These results portray BHL to be a well distributed biodiversity collection across all kingdoms of life. In order to target areas for future digitization in BHL, further analysis into lower levels of taxonomic rank should be executed. This process can be done in the same manner as described above with minor manipulation to the name_analysis.py.[21] These visualizations should be provided for public use as well as internal use so that users can browse the collection by taxonomic rank.

Topical Analysis

BHL metadata contains “subject” keywords from MARC records. These subjects are from Library of Congress Subject Headings (LCSH) and can give information about an item’s topic, genre, or format. Almost all items in BHL (97%) have at least one subject keyword assigned to it.[22],[23] Like BHL’s scientific names, this is a list of words that needs to be organized or set against a schema in order to make them meaningful as descriptors of the full collection.

I began researching LCSH to see if there was a way to map BHL’s subject keywords back to the authority files and discovered the work of Dr. Charles Julien at McGill University. Dr. Julien has been working on turning the LCSH into a hierarchical tree for browsing purposes[24],[25].

---

[21]Name_analysis.py available at https://osf.io/xnu8b/
[22]Hannah Tarver, et al. *An Exploratory Analysis of Subject Metadata in the Digital Public Library of America*. (International Conference on Dublin Core and Metadata Applications, 2015).
[23]Corey A. Harper. *Metadata Analytics, Visualization, and Optimization: Experiments in statistical analysis of the Digital Public Library of America (DPLA)*. (Code 4 Lib, 2016).
[24]Charles-Antoine Julien, et al. *Constructing a True LCSH Tree of a Science and Engineering Collection*. (Journal of the American Society for Information Science and Technology, 2012).

LCSH are not hierarchical, and in order to create a hierarchy, terms must all come from one base node and all cyclical relationships must be eliminated. This hierarchy was created for science terms and was tested with bibliographic records from the science library at McGill where records were matched to the hierarchy to create a browsable tree of the collection.

I contacted Dr. Julien to attempt to use the same program for the records from BHL. Unfortunately, the program required MARC formatted LCSH authority files and Library of Congress only makes authority files available as MADS/RDF and SKOS/RDF. BHL technical developer, Mike Lichtenberg, was able to turn SKOS into MADS and then turn MADS into MARC. He also wrote an SQL script to extract bibliographic information from BHL ingest data in the format expected by the application. However, the reformatted MARC files did not work correctly with the application. This project was therefore dropped on June 21, 2017.

I then began to think about creating new subject keywords directly from BHL text using text mining methods. At DPLAFest in April 2017, JSTOR Labs gave a presentation about Topicgraph, an open-source topic-based navigational aid. This tool uses an in-house thesaurus and topic modeling to extract topics from PDFs. The beta version allows you to upload your own PDFs to analyze, but no options currently exist for analyzing a bulk amount of literature. I reached out to the developers at JSTOR Labs to explain my NDSR project with BHL. We began an exploratory study and I provided a random sample of URLs to 100 BHL texts to get started.[26] I also met with their metadata librarian to determine if their in-house thesaurus would capture the types of subject areas most important to BHL. The thesaurus draws from International Plant Names Index, NASA, and other established vocabularies that appeared relevant to biodiversity literature terms. After the researchers at JSTOR Labs were able to successfully set up a workflow with BHL content, I sent a larger random sample of 10,000 URLs from BHL. From this sample, 8952 records, or 90% of the random sample, matched to at least one topic from JSTOR's thesaurus.[27] The developers were able to make a drill down topic map of the BHL items and a radar map that plots BHL's subset against JSTOR's collection.[28]

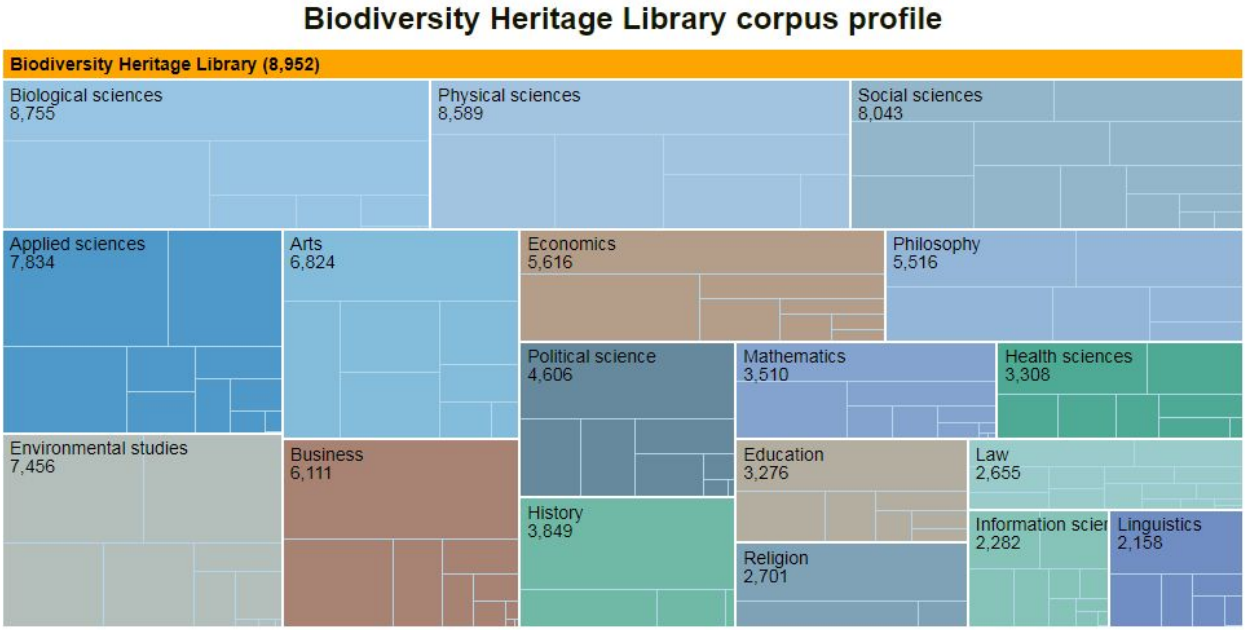
---

[25]Charles-Antoine Julien, et al. *Capitalizing on Information Organization and Information Visualization for the Future Generation Catalogue.* (Library Trends, 2012).

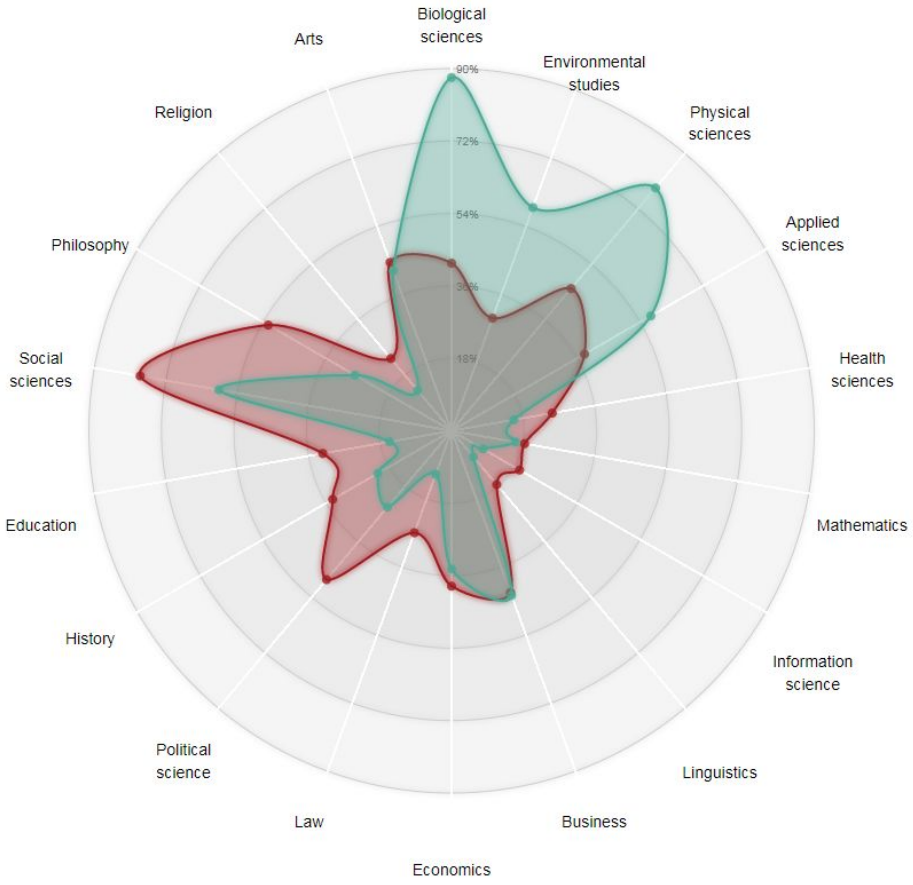
[26] Steps for creating a random sample from BHL https://osf.io/5zgh6/

[27] Raw results available in spreadsheet format here https://osf.io/5jz7s/

[28]Active graphs at https://jstor-labs.github.io/corpora/?corpus=bhl,jstor (not to be shared publicly).



This graph groups found topics from BHL into JSTOR's established hierarchy of topics. The top level of the hierarchy is made up of 18 terms, all of which were represented in the BHL sample. Texts can contain more than one topic, which is why the sum of the totals for each topic is greater than the original sample size. It is not surprising that biological sciences was determined to be a topic in 8,755 of the 10,000 texts from the BHL sample. This drill down map goes further into the topic hierarchy as you click through topics. Developers at JSTOR Labs say it is possible to add links back to the text from this visualization, potentially making this a browsing tool for users.



The radar map created by JSTOR Labs plots BHL's collection, in teal, against JSTOR's collection, in red. On average, JSTOR's collection skews towards social sciences with about an equal representation of political science, economics, business, arts and physical sciences. BHL on the other hand skews greatly towards biological sciences and physical sciences with a small representation of environmental studies and applied sciences.

These results received positive feedback from the Collections Committee. I encourage continuing the collaboration between BHL and JSTOR Labs especially in the work of topic analysis. The next step for BHL is to determine the purposes for this type of analysis in order to create a focused plan for JSTOR Labs to execute. These types of visualizations can be used both as a browsing tool for users and as a development tool for the Collections Committee to determine which areas to focus future digitization efforts.

Geographic Analysis

Currently, the only metadata available that gives any geographic representation to BHL items is in the publication information. This however, does not represent the geography of the content of the item, just the place where it was published. In order to represent the geographies covered in the text, the text must be mined for place names.

Geoparsing is the process of pulling out toponyms, or place names, out of a text. After extracting toponyms, a gazetteer can be used to assign longitude and latitude coordinates to place names. These coordinates can then be plotted onto a map to visualize the geographic distribution of text coverage. This whole process can be done in a modular way, by choosing different tools to perform each step, or by use of a pipeline tool which performs all of the steps in one process.

I researched several modular tools and tested a couple pipeline tools with texts from BHL in efforts to develop a successful workflow. There are still many common pitfalls of geoparsing text, including the ability to decipher context and tell the difference between Athens, Greece and Athens, Texas, for example, but this should not discourage BHL from performing named entity extraction for geographic names.[29]

To first identify place names from the text, I found Geograpy, a Python tool that extracts countries, regions and cities from URLs or blocks of text.[30] Geograpy uses NLTK (natural language tool kit), a Python platform for working with human language. Geograpy currently only works for English and I was not able to properly install the tool and was not able to test it. Geopy is a Python library that locates coordinates of place names.[31] Since I was not able to test Geograpy to extract names from BHL, I did not use Geopy to locate place name coordinates. Rather, I was able to test multiple all-in-one pipeline geoparsers.

Geoparser.io is a web API, built on RESTful web services, an architecture that uses a uniform and predefined set of stateless operations such as HTTP, that identifies place names mentioned in text, disambiguates those names, and place names found in the text in JSON file format.[32] Calls are limited to 10,000 characters so files have to be read in line by line. This amount of calls quickly surpasses the free version and even the “Developer” version which allows for 100,000 calls per month for $1000/month. This is not a feasible tool for a corpus the size of BHL.

The Edinburgh Geoparser is “a language processing tool designed to detect placename references in English text and ground them against an authoritative gazetteer so that they can be plotted on a map.”[33] However, according to its documentation, the Edinburgh Geoparser does not work well for large documents made up of sub-sections, like journals and book chapters and therefore does not seem like a good fit for BHL.

Argo is “a workbench for building and running text-analysis solutions.”[34] This tool was used in the Mining Biodiversity project with BHL materials in 2016/2017. I contacted Mining Biodiversity project members, William Ulate and Riza Batista-Navarro, to use workflows from Mining Biodiversity project. Dr. Batista-Navarro shared workflows on 5/15 for recognizing location names, but warned that it was used on the “ACE 2005 corpus,” a collection of news-related documents from various spoken and written sources, and therefore may not have satisfactory results for biodiversity materials. I tested this workflow on BHL materials and did not

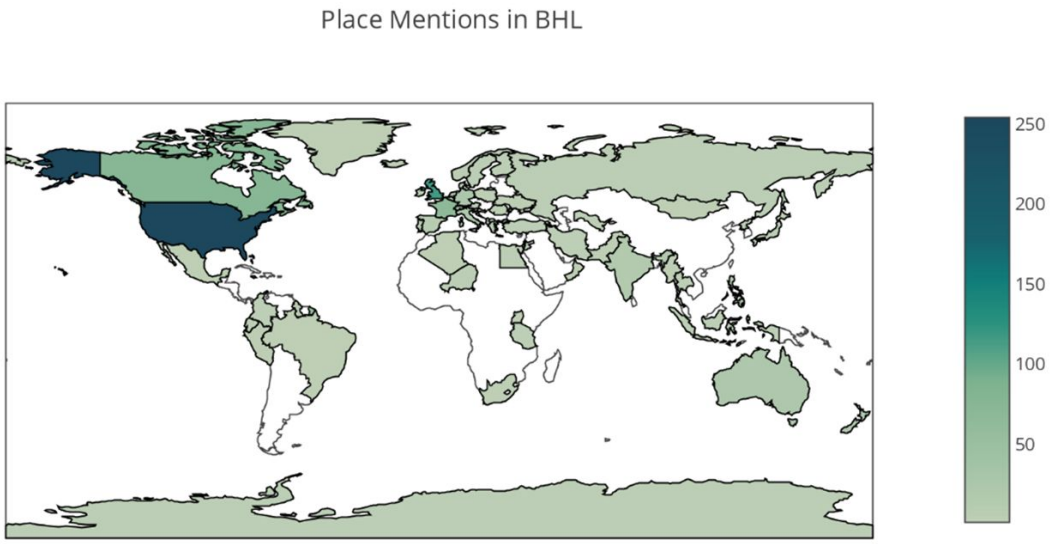
---

[29] Milan Gritta, et al. *What’s missing in geographical parsing?* (Lang Resources & Evaluation, 2017).
[30] https://pypi.python.org/pypi/geograpy
[31] https://geopy.readthedocs.io/en/1.11.0/
[32] https://geoparser.io/
[33] http://groups.inf.ed.ac.uk/geoparser/documentation/v1.1/epub/index.html
[34] http://argo.nactem.ac.uk/

get expected results, more customization of the Mining Biodiversity workflows is needed for Argo work well with BHL text.

Geoparsepy is “a Python geoparsing library that will extract and disambiguate locations from text.[35] It uses a local OpenStreetMap database tables which allows a very high geoparsing throughput and no rate limits.” Unfortunately, Geoparsepy proved to have a difficult set up and installation and I was not able to run any texts through this tool.

The pipeline tool I had the most success with has been MediaMeter’s CLIFF. CLIFF is “a piece of software for extracting geolocation data from text, bundled into a server that can be accessed via API calls.”[36] I set CLIFF up in a virtual machine following the instructions on the CLIFF Github page.[37] I was able to successfully get results using BHL texts, however, the processing time is quite slow especially for the size of the collection. The results from CLIFF are country codes from the Geonames gazetteer. These country codes can be used as the data points for creating a map. To create a map with the CLIFF results, I used plot.ly, an online data visualization and analytics tool.[38]



While this graph represents a small subset from the BHL collection, it makes a case for continuing this type of work with a larger amount of text. Steps need to be taken to decrease the processing time of CLIFF in order to create a more representative image of the BHL collection. This geographic information can potentially be crossed with BHL scientific names and occurrence records from other sites to track species across space. Visuals created with geographic information can be used by users as a browsing tool for the collection. And statistics on geographic coverage can help target areas for further digitization in BHL.

---

[35] https://pythonhosted.org/geoparsepy/index.html
[36] http://cliff.mediameter.org/ or https://cliff.mediacloud.org/
[37] https://github.com/mitmedialab/CLIFF-up
[38] https://plot.ly/create/?fid=alesquivel:1

I recommend doing additional geoparsing of the BHL corpus with CLIFF with a bigger set of texts and eventually resolving terms to a more granular geography than country.

Statistical Analysis

In order to determine how much literature BHL has not yet digitized, the scope of biodiversity literature must be calculated. To estimate the scope of biodiversity literature, I used an approach from population ecology called capture-mark-recapture (CMR) that has in recent years been adapted for bibliometric analysis.

Capture-mark-recapture is a statistical approach to quantifying the unknown size of a population.[39] For example, in order to determine the population size of perch in a lake, you can go out and take one sample of perch and mark them all in a distinguishing manner. You release the marked sample back into the wild to mix back in with the full population. You then take another sample of perch, and can assume that the proportion of marked perch in the second sample represents the proportion of the first sample to the whole population. Represented as an equation, when N equals the total population, ng equals the first capture, nm equals the second capture, and n0 equals the recaptured.

$$\frac{n0}{nm} = \frac{ng}{N}$$

This technique has been adapted and used in bibliometrics to determine the amount of literature published in different subject areas. Several studies apply different statistical models (such as Lincoln-Peterson model, Poisson regression model, and Seber probabilistic model), but use the same general CMR principles.[40],[41],[42],[43]

Given this existing research, I completed a case study for BHL using a subset of biodiversity literature. Essentially, I performed multiple captures for bibliographic records in different areas for the same subset of biodiversity literature. The overlap between results could be extrapolated to understand the total size of the literature subject in question. The selected subset I used was the genus artemisia, which has between 200-400 species including sagebrush, mugwort, and wormwood. This genus of plants spans a large geographic area and poses challenges for this type of work. For example, a keyword search for “artemisia” might return results for the Italian painter from the Baroque period, Artemisia Gentileschi, which would be outside of the scope of BHL.

---

[39]William J. Sutherland. *Ecological Census Techniques: a Handbook*. (Cambridge University Press, 1996).
[40]M. Kastner, et al. *The Capture-mark-recapture Technique Can Be Used as a Stopping Rule When Searching in Systematic Reviews*. (Journal of Clinical Epidemiology, 2009).
[41]D. Lane, et al. *Capture-mark-recapture as a tool for estimating the number of articles available for systematic reviews in critical care medicine.* (Journal of Critical Care, 2013).
[42]Madian Khabsa and C.Lee Giles. *The Number of Scholarly Documents on the Public Web.*(PLOS ONE, 2014).
[43]Arturo H. Ariño. *Approaches to estimating the universe of natural history collections data.* (Biodiversity Informatics, 2014).

For my project I performed three captures for artemisia for the time period 1950-1980, one in WorldCat, one in Google Scholar, and one in BHL. The query for WorldCat and Google Scholar were identical and included common names for species of artemisia and excluded non-biodiversity terms just as “painting” and “recipe.” The results from WorldCat (1732 records) were exported to EndNote and then exported out of EndNote. I manually filtered some results from WorldCat by excluding titles that were obviously out of scope for BHL.

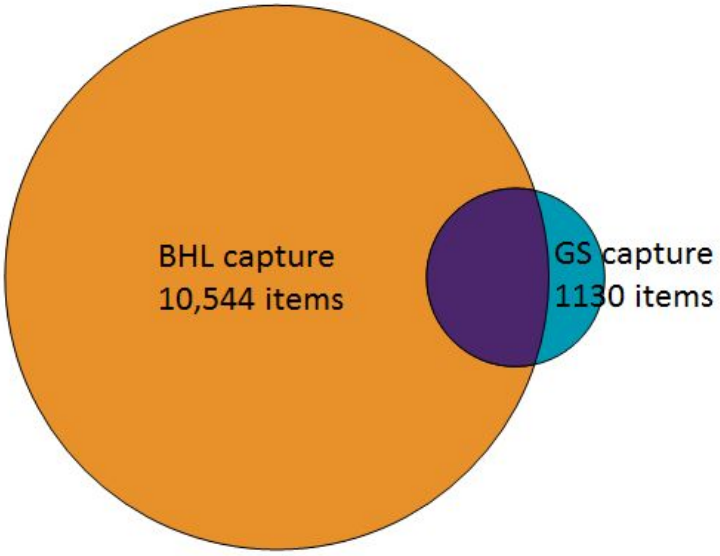
The Google Scholar results proved more challenging. I was able to submit the same query used for WorldCat in Google Scholar but there currently is no easy way to export the results. Google’s code of conduct discourages web scraping results and doing so can get your IP address blocked. Because of this restriction I used a tool called Publish or Perish which gathers 1000 results from Google Scholar with the highest citation metrics.[44] I was then able to export these results into a csv file. After reviewing initial Google Scholar results from Publish or Perish I noticed a lot of out of scope titles and ran a second query omitting the keyword artemisinin. I then compiled both lists of results, removed duplicates and ended up with a total of 1256 results from Google Scholar.

For the control capture, I used titles from BHL. Since scientific names are indexed in BHL I was able to query the BHL API using a query written in R to collect a list of all titles that have a species of artemisia in them.[45] This first required getting a list of all the species names in artemisia from BHL (1500) and then querying each of those names to get titles which resulted in 10,544 records.

After performing these three captures, I compared the results to each other to determine the ratio of overlap between BHL and Google Scholar and BHL and WorldCat. This was done in Python by comparing titles from each capture to each other.[46] First, any characters that were not alphanumeric were removed from title strings and all characters were set to lowercase. Then, strings were truncated to the shorter string when compared to each other and fuzzy matching was used when making the comparisons. Fuzzy matching allows for strings to be considered matched to each other with less than 100% accuracy. This is done by manipulating the variable known as the edit distance. The edit distance is how many changes need to occur before the strings match. I compared the capture results to each other using an edit distance of 0, 1, 2, and 3. The higher the edit distance, the more matches are found. After determining the amount of matches between results, the total amount of published artemisia literature can be determined.

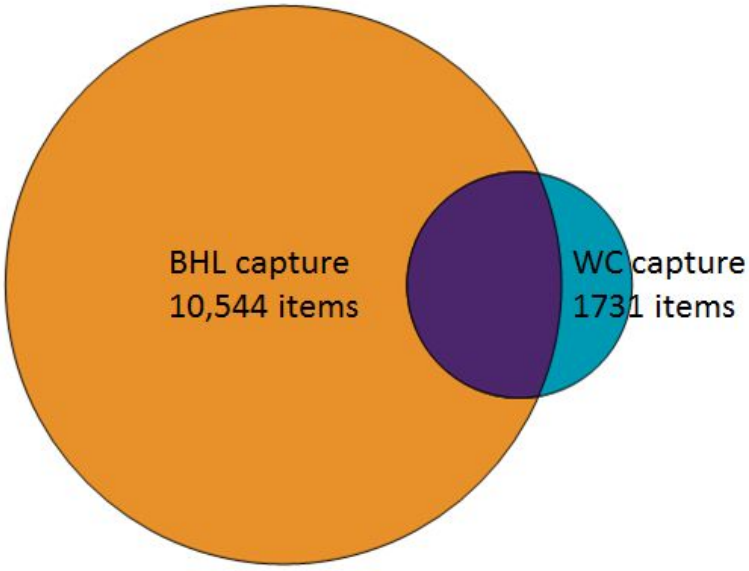
[44] https://harzing.com/resources/publish-or-perish
[45] Query in R available here https://osf.io/c73r6/
[46] Python script available here https://osf.io/dqetb/

BHL and Google Scholar CMR, edit distance 3.



$$\frac{n0}{nm} = \frac{ng}{N} \qquad \frac{792}{1130} = \frac{10{,}544}{15{,}043}$$

BHL and WorldCat CMR, edit distance 3.



$$\frac{n0}{nm} = \frac{ng}{N} \qquad \frac{1199}{1731} = \frac{10{,}544}{15{,}222}$$

The results of these CMRs put BHL's coverage of published works on the genus artemisia from 1950-1980 at 70% according to Google Scholar capture and 69% according to WorldCat capture. However, it must be noted that these results are affected by not having a full capture of Google Scholar and are calculated with the most basic application of CMR.

This case study lays out the framework for performing statistical analysis for determining the scope of a subject of literature. BHL can redo this process using another sample. This can also be adopted by other digital libraries to determine sizes of other subject areas.

Recommendations

In conclusion, I have several recommendations for BHL to address issues of gaps in content. First, I recommend using temporal analysis to target missing journal issues in the collection. According to the user survey results from fellow NDSR resident Pam McClanahan, BHL users that notice gaps in content specifically notice gaps in journal runs.

Next, I suggest filtering the scientific names found in BHL into further taxonomic ranks. Further filtering can give a more granular look into the collection to address areas with lack of representation in the literature. I also suggest that this information and or visualizations be made available to users as a browsing tool.

I propose a continued collaboration with JSTOR Labs to create more topical analyses of BHL literature.

I advise geoparsing more BHL texts with CLIFF after consulting a software engineer about how to decrease processing time for the application. The geoparsed names from BHL can eventually be used as named entities that add an additional entry point into the BHL collection.

Lastly, I suggest doing another statistical analysis of biodiversity literature with a different subject area.

Role at Lenhardt Library

The host institution plays a large role in this residency program. In addition to my project work, I was able to gain experience working at a special library. While hosted at the Chicago Botanic Garden's Lenhardt Library, I attended numerous trainings, worked library special events, and carried out a digitization project for the library.

As a result of my payroll status, I was categorized as a volunteer at the garden and therefore attended the volunteer training. This training session gave me an overview of the garden history and helped me understand the huge role volunteers fill at the garden. At the library, my training included shadowing all library staff to understand each person's functions and responsibilities at the library. I was trained on the circulation desk at the library and was able to cover circulation shifts when needed. Working at the circulation desk gave me the opportunity to interact with library patrons and to become more familiar with the library holdings. A few months into the residency, my mentor arranged for me to meet with the head of Living Plant Documentation department to learn about data management at a botanic garden. And as

necessary, I attended special training sessions at the garden such as a presentation on boxwood blight.

The Lenhardt Library hosted several public events during my residency that I participated it. The garden celebrated World Environment Day and I was at the library giving live demonstrations of BHL. For Carl Linnaeus' birthday, the library showcased some of its rare books and I helped monitor the room during that event.

At the beginning of my residency I began a small project for the library in which I assessed a specific collection, the Massachusetts Horticultural Society (MHS) collection, for possible digitization. This collection was purchased by the library from the Massachusetts Horticultural Society in 2002 and contains horticultural and gardening journals. I used a list in Microsoft Access to look up journal titles in BHL to determine if any missing issues could be filled with issues from the MHS collection. If I found a missing journal issue in BHL that could be filled with material from the MHS collection, I then searched in WorldCat to determine the rarity of the journal issue in question. This information could then be used by library director Leora Siegel to choose candidates for digitization.

Additionally, as a member of the library staff, I attended bimonthly library staff meetings and biannual garden staff meetings. I also wrote a blog post for the garden's blog about my residency and was interviewed for the garden's member magazine, Keep Growing.

Conferences

Presenting an aspect of work at a professional conference was one of the deliverables for this grant. My fellow residents and I were able to present and attend multiple conferences this year. In April, we presented as a cohort with mentors, Leora Siegel and Trish Rose-Sandler, at DPLAFest in Chicago. DPLAFest is hosted by Digital Public Libraries of America (DPLA), a network of libraries, archives, museums and cultural institutions that operates on a "hub" model. BHL is a content hub for DPLA which means BHL content gets passed along to DPLA and can be accessed from their portal. Our group presentation gave an overview of NDSR, BHL, and each of our individual projects and the mentor-mentee relationship.

The next week, the residents attended and presented in Washington D.C. at the NDSR Symposium. This inaugural symposium brought together past, present, and future NDSR residents, mentors, and hosts to analyze and discuss past projects and work towards improvements of the program for the future. Our cohort held a panel discussion about communication in a geographically dispersed cohort. In the history of NDSR programs, we are only the second cohort that is geographically dispersed.

In June, I presented at the inaugural Digital Data in Biodiversity Research at University of Michigan in Ann Arbor. This conference brought together researchers, librarians, and collections managers in the biodiversity field to discuss working with digital data. I gave a presentation about my performing statistical analysis using capture mark recapture to calculate the size of biodiversity literature and created a poster about how to visualize a biodiversity literature collection.

Later in June, Ariadne Rehbein and I represented the cohort at ALA Annual Conference and Exhibition in Chicago. This conference brings together professionals from a wide range of libraries from all over the world. The cohort created a poster that gave midyear updates on each of our projects' progress.

In September I presented a poster at iPres, the international conference on digital preservation in Kyoto, Japan. I shared preliminary visualizations from performing a content analysis on a digital library of biodiversity literature.

And in October I attended the Midwest Data Librarians Symposium hosted by Purdue University. This regional symposium focused on different aspects of data librarianship including cyber security, data visualization, and data management.

I'm grateful to be given the opportunity to attend so many professional conferences during this residency program.

Appendix

Bibliography

Ariño, Arturo H. “Approaches to Estimating the Universe of Natural History Collections Data.” *Biodiversity Informatics*, 7 (2010): 81–92.

Ariño, Arturo H., Vishwas Chavan and Javier Otegui. *Best Practice Guide for Data Gap Analysis for Biodiversity Stakeholders.* (GBIF, 2016).

Gritta, Milan, Pilehvar, M.T., Limsopatham, N., Collier, N. “What’s Missing in Geographical Parsing?” *Language Resources and Evaluation,* (2017): 1-21.

Harper, Corey A. “Metadata Analytics, Visualization, and Optimization: Experiments in Statistical Analysis of the Digital Public Library of America (DPLA)” *Code 4 Lib.,*33 (2016).

Julien, Charles-Antoine,Tirilly, P.,Leide, J.E., and Guastavino, C. “Constructing a True LCSH Tree of a Science and Engineering Collection.” *Journal of the American Society for Information Science and Technology,* 63:12 (2012): 2405-2418.

Julien, C.A., Bouthillier, F., and Guastavino, C. “Capitalizing on Information Organization and Information Visualization for the Future Generation Catalogue.” *Library Trends*, 61(1) (2012): 149-161.

Kastner, Monika, Straus, S.E., McKibbon, K. A., and Goldsmith, C.H. "The Capture-mark-recapture Technique Can Be Used as a Stopping Rule When Searching in Systematic Reviews." *Journal of Clinical Epidemiology,* 62:2 (2009): 149-157.

Khabsa, Madian and C. Lee Giles. “The Number of Scholarly Documents on the Public Web.” *PLOS ONE,* 9:5 (2014).

Lane, Daniel, Dykeman, J., Ferri, M., Goldsmith, C., Stelfox, H. (2013). “Capture-mark-recapture as a Tool for Estimating the Number of Articles Available for Systematic Reviews in Critical Care Medicine.” *Journal of Critical Care,* 28 (2013): 469-475.

Larsen, Brendan B., Miller, E.C., Rhodes, M.K., and Wiens, J.J. “Inordinate Fondness Multiplied and Redistributed: the Number of Species on Earth and the New Pie of Life.” *The Quarterly Review of Biology*, 92:3 (2017): 229–265.

Patterson, David, Mozzherin, D., Shorthouse, D., Thessen, A. “Challenges with Using Names to Link Digital Biodiversity Information.” *Biodiversity Data Journal,* 4 (2016).

Sutherland, William J. *Ecological Census Techniques: a Handbook.* Cambridge University Press, 1996.

Tarver, Hannah, Phillips, M., Zavalina, O., Kizhakkethil, P."An Exploratory Analysis of Subject Metadata in the Digital Public Library of America." *International Conference on Dublin Core and Metadata Applications*, 2015.

Technical documents

Temporal Technical Document

Guide to analyzing the BHL collection by year and producing a histogram.

1. Download files from BHL export page http://biodivlib.wikispaces.com/Data+Exports
   1.1. title.txt and item.txt
2. Filter selected columns. Keep only title.titleid, title.startyear, title.endyear, title.callnumber and item.titleid, item.itemid, item.year and item.volumeinfo. (Not mandatory, but the code works with these columns).
3. Merge tables in Python by titleid. Each record now has a field for year, startyear, and endyear.
4. Select a year for each item.
   4.1. If year field is valid, use that date for the record.
   4.2. If year field is not valid, and startyear and endyear are valid, take the range between startyear and endyear.
   4.3. Distribute the item across the range. For example, a range of 5 years results in ⅕ of an item counted towards each year in range.
5. Use matplotlib library in Python to create histogram and save as png.

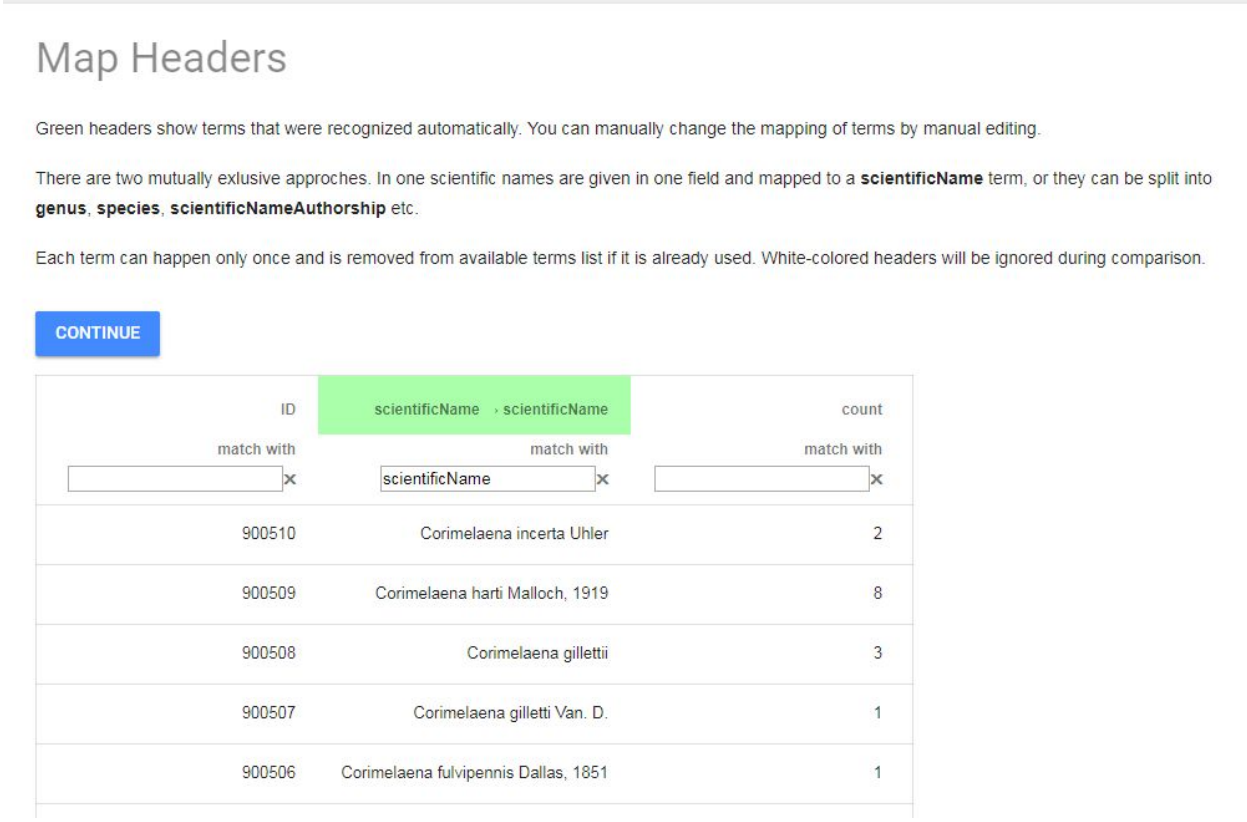
Steps 3-5 can be replicated using the Python code found at https://osf.io/3yuhr/

Taxonomic Technical Document

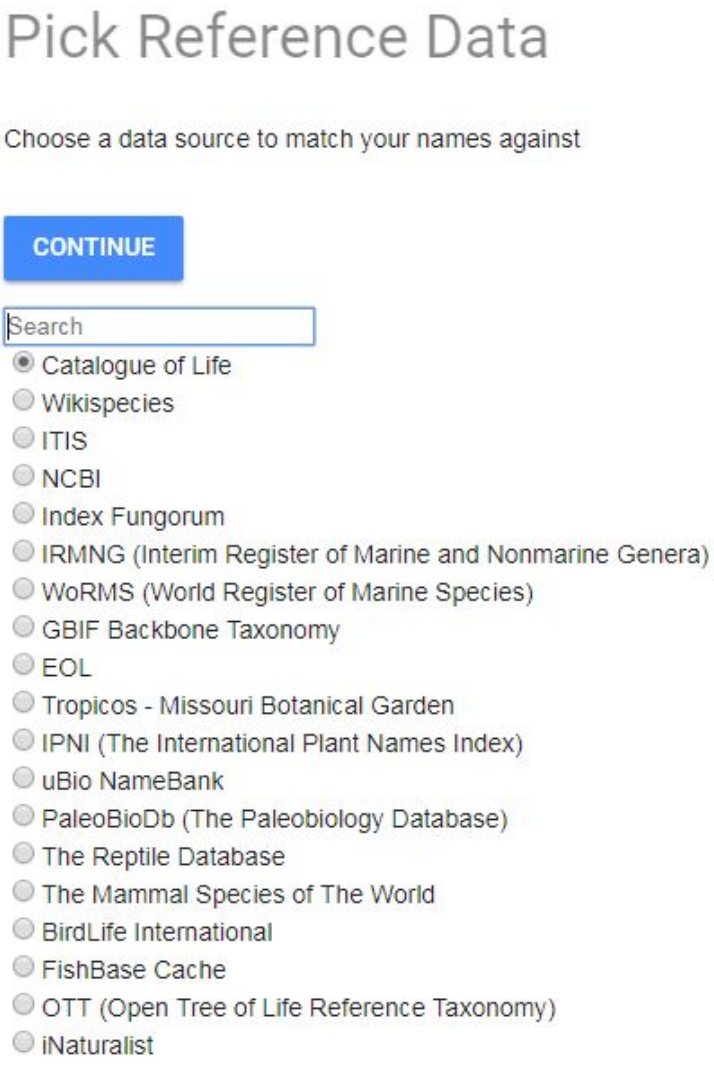
Guide to analyzing scientific names in BHL and filtering into buckets by kingdom.

1. Download list of scientific names from BHL export page (zip file of all tables for pagename table) http://biodivlib.wikispaces.com/Data+Exports
2. Use Python to create a dictionary that counts all names and gives a list of unique names. Python code available here https://osf.io/s7fdw/
3. Split dictionary into smaller lists of 100,000 names per file using Python. Python code available here https://osf.io/g73af/
4. Use Global Names List Resolver and upload one file of 100,000 names http://listresolver.globalnames.org/
   4.1. Make sure headings match. (For this example, our input file has a page number ID from BHL, scientific name from BHL, and count from the Python dictionary.

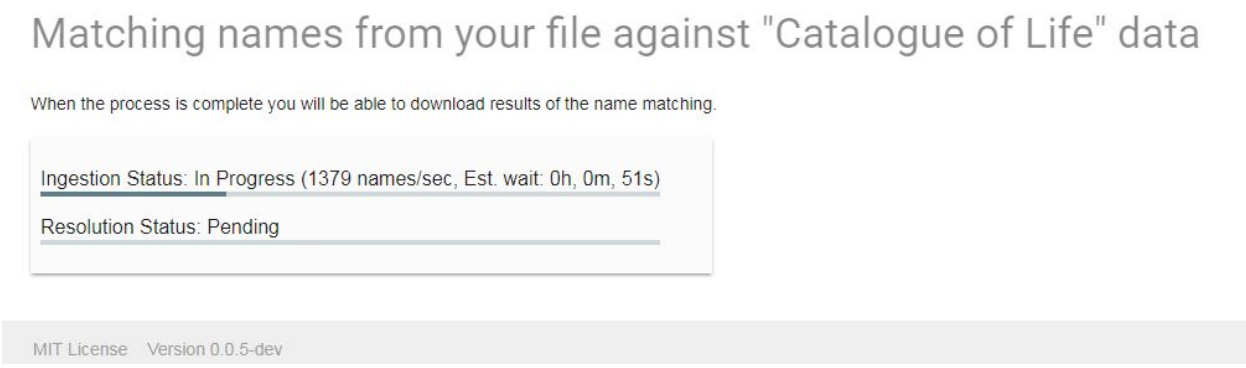
Only scientific name header needs to match to scientificName header in Global Names).



4.2. Select a data source to compare names to (Catalogue of Life was used in this example).



4.3. Click continue and wait for results.



4.4. Export results as CSV file.

5. Analyze results. All of step 5 can be completed using this Python script https://osf.io/xnu8b/

5.1. Using Python, drop all unmatched names.
5.2. Extract kingdom from classification paths.
5.3. Remove duplicates (some input values, scientific names, matched to multiple output values).
5.4. Count up kingdoms.
5.4.1. Multiply kingdom occurrences by the count.

6. Visualize the results
6.1. Upload results into Excel and created a clustered column chart.
6.1.1. Set y-axis to base 10 log scale.
6.2. Add species estimates for each kingdom from data source.

Geographic Technical Document

Guide for geoparsing OCRd text from BHL and creating a heatmap of the results.

1. Install virtual machine
1.1. Follow these instructions https://github.com/mitmedialab/CLIFF-up
2. Download full text in bulk from Internet Archive
2.1. Download collections from Internet Archive and select which file types to download. Use this script to download the djvu files of a collection. https://osf.io/aq7kg/
2.1.1. Adapted from the instructions found here https://programminghistorian.org/lessons/data-mining-the-internet-archive
3. Send files through CLIFF for analysis. This can be done using Python code found here https://osf.io/bm2sv/
3.1. Loop files through CLIFF from one folder.
3.2. Get country codes from CLIFF and create dictionary of results
3.3. Convert country codes to ISO alpha 3 (required for plot.ly chart)
3.4. Write results to CSV file.
4. Plot results into plot.ly to create heat map.
4.1. Visit https://plot.ly/ and select “Create” > “Chart”
4.2. Select the “choropleth” chart.
4.3. Import data file or copy and paste data into grid fields.
4.4. Make sure you are using correct country code format, ISO alpha 3

Statistical Analysis Technical Document

Guide to using capture-mark-recapture to calculate the unknown size of a subject of literature.

1. Select organism or subject area from biodiversity field
1.1. Select appropriate/relevant time period for given organism
2. Create queries that include common names of organism, including common misspellings, and exclude keywords that may result in out of scope results.[47]
3. Query selected sources

[47] Example of query from artemisia CMR available here https://osf.io/n9r63/

3.1. WorldCat
    3.1.1. Enter query into WorldCat.
    3.1.2. Export results as EndNote records then upload to EndNote.
    3.1.3. Export from EndNote as RIS.
    3.1.4. Upload RIS to Zotero.
    3.1.5. Download from Zotero as CSV file.[48]
3.2. Google Scholar
    3.2.1. Use Publish or Perish tool to grab records with highest citation metrics.[49]
    3.2.2. Select Google Scholar as search engine.
    3.2.3. Enter same query as used in WorldCat
    3.2.4. Export results as CSV
3.3. BHL
    3.3.1. Obtain API key from BHL https://www.biodiversitylibrary.org/getapikey.aspx
    3.3.2. Write query in R to pull all items with certain scientific name.[50]
    3.3.3. Filter down results by removing duplicate titleIDs.

4. Manually clean results
    4.1. Open in each capture in Excel
    4.2. Remove things that are obviously out of scope.
5. Compare results to each other.[51]
    5.1. Compare BHL titles to WorldCat results. Compare BHL titles to Google Scholar resutls.
    5.2. Turn all title strings into alphanumeric characters only.
    5.3. Turn all title strings into lowercase characters.
    5.4. When comparing title strings, truncate all title strings to the shorter title string.
    5.5. Set edit distance.
6. Extrapolate total population size using ratio of capture overlap.
    6.1. $\frac{n0}{nm} = \frac{ng}{N}$
    6.2. N equals the total population, ng equals the first capture, nm equals the second capture, and n0 equals the recaptured.
    6.3. Plug in results to estimate size of total population.

[48] Export steps are necessary because we need records in csv format. Other workarounds are possible.
[49] Download and install Publish or Perish from here https://harzing.com/resources/publish-or-perish
[50] R query for artemisia found here https://osf.io/c73r6/
[51] Python script for comparing results found here https://osf.io/dqetb/

List of Posters and Presentations

All presentations and posters can be found at https://osf.io/urkcn/files/

DPLAFest, Chicago, IL. Presentation. *Foundations to Actions: Extending Innovations to Digital Libraries in Partnership with NDSR Learners.* Esquivel, A., Kings, M., McClanahan, P., Mika, K., Rehbein A., Rose-Sandler, T., & Siegel, L., (April 2017).

NDSR Symposium, Washington, D.C. Presentation. *NDSR at BHL Building a Communication Network for Collaborative Projects.* Esquivel, A., Kings, M., McClanahan, P., Mika, K., & Rehbein A. (April 2017).

Inaugural Digital Data in Biodiversity Research Conference, Ann Arbor, MI. Presentation. *Using Statistical Analysis to Calculate the Size of Biodiversity Literature.* Esquivel, Alicia, (June 2017).

Inaugural Digital Data in Biodiversity Research Conference, Ann Arbor, MI. Poster. *Visually Representing a Biodiversity Literature Collection Using Metadata and Full Text Analysis.* Esquivel, Alicia, (June 2017).

American Library Association Annual Meeting, Chicago, IL. Poster. *Improving Access to the Biodiversity Heritage Library: Halfway Remarks by the National Digital Stewardship Residents.* Esquivel, A., Kings, M., McClanahan, P., Mika, K., & Rehbein A. (June 2017).

Technical Team Meeting, St. Louis, MO. Presentation. *Content Analysis.* Esquivel, Alicia. (September 2017).

iPres: International Conference on Digital Preservation, Kyoto, Japan. Poster. *Performing a Content Analysis of Biodiversity Literature.* Esquivel, Alicia. (September 2017).

Content Analysis. Webinar. Esquivel, Alicia. (November 2017). http://idigbio.adobeconnect.com/pcb8q6y4yy16/

Links to blog posts

From Resident blog, authored by me:

Defining the Scope of Biodiversity Literature 2/23/2017
https://ndsrbhl.wordpress.com/2017/02/23/defining-the-scope-of-biodiversity-literature/
Data Management at a Botanic Garden 4/24/2017
https://ndsrbhl.wordpress.com/2017/04/24/data-management-at-a-botanic-garden/
Digital Data in Biodiversity Research 6/14/2017
https://ndsrbhl.wordpress.com/2017/06/14/digital-data-in-biodiversity-research/
NDSR at iPres 2017 10/12/2017
https://ndsrbhl.wordpress.com/2017/10/12/ndsr-at-ipres-2017/

From Resident blog, collaboratively authored by all Residents:

BHL Bootcamp 3/3/2017
https://ndsrbhl.wordpress.com/2017/03/03/bhl-bootcamp/
DPLAFest and NDSR Symposium 5/4/2017
https://ndsrbhl.wordpress.com/2017/05/04/dplafest-and-ndsr-symposium/
BHL NDSR at ALA 7/7/2017
https://ndsrbhl.wordpress.com/2017/07/07/bhl-ndsr-at-ala/
Preliminary Recommendations from NDSR Residents 9/22/2017
https://ndsrbhl.wordpress.com/2017/09/22/preliminary-recommendations-from-ndsr-residents/

From BHL blog, collaboratively authored by all Residents:

NDSR Residents Update 9/19/2017
http://blog.biodiversitylibrary.org/2017/09/ndsr-residents-update.html
NDSR Residents Mid Year Update 7/18/2017
http://blog.biodiversitylibrary.org/2017/07/ndsr-residents-mid-year-update.html
An Update from NDSR Residents 5/16/2017
http://blog.biodiversitylibrary.org/2017/05/an-update-from-ndsr-residents.html
Keeping Up with NDSR 3/21/2017
http://blog.biodiversitylibrary.org/2017/03/keeping-up-with-ndsr.html

From CBG Library blog, authored by me:

Documenting and Digitizing Biodiversity, 5/9/2017
http://my.chicagobotanic.org/education/lenhardt-library/documenting-and-digitizing-biodiversity/