Performing a Content Analysis of Biodiversity Literature

Alicia Esquivel
National Digital Stewardship Resident
Lenhardt Library
Chicago Botanic Garden
esquivelndsr@gmail.com

ABSTRACT

Biodiversity essentially describes all living organisms and their environments. As the world's most comprehensive digital library collection of legacy biodiversity literature, the Biodiversity Heritage Library (BHL) is dedicated to building and maintaining their repository by determining the scope of biodiversity literature and analyzing the collection strengths, weaknesses and gaps in the BHL corpus [1]. Regular analyses have been performed since the creation of BHL in 2006 and as the corpus expands and methods for analyzing literature in the public domain develop, continuous analysis of BHL coverage and gaps remains critical. In the course of 2017, BHL has the opportunity to work with a team of National Digital Stewardship Residents who are committed to developing a plan for the next generation of BHL.

This poster will illustrate the current research and progress of the collection analysis project of the National Digital Stewardship Resident (NDSR) project led by the resident hosted at the Chicago Botanic Garden's Lenhardt Library.

KEYWORDS

National Digital Stewardship Residency, NDSR, Biodiversity Heritage Library, BHL, biodiversity, collection analysis, content analysis, digital library

1 INTRODUCTION

1.1 Biodiversity Heritage Library

The Biodiversity Heritage Library (BHL) is an open access digital library for biodiversity literature. Member and affiliate partners of BHL form a consortium of natural history and botanical libraries that work together to digitize legacy literature on biodiversity held in their collections to be used as a global "biodiversity commons." As of March 2017 over 51,000,000 pages have been digitized and made available since digitization began in 2006. In addition to public domain books and journals, BHL works to obtain permission from publishers to digitize biodiversity materials still under copyright. BHL expanded globally in 2009 and now has nodes in Europe, China, Australia, Brazil, Egypt, Africa, and Singapore [2]. In June 2016 BHL was awarded a grant funded by the Institute of Museum and Library Services (IMLS) to work with 5 residents hosted at BHL member and affiliate institutions across the United States. The "Foundations to Actions: Extending Innovation in Digital Libraries in Partnership with NDSR Learners" (Foundations to Actions) hosts residents for 12 months over the course of 2017.

1.2 National Digital Stewardship Residency

The National Digital Stewardship Residency (NDSR) program aims to develop a community of professionals in the fields of digital stewardship and informatics through collaborative field experience. Since the 2013 pilot program, NDSR has provided opportunities for recent graduates of library and information science to link theoretical knowledge to practice in a professional context through hands-on experiences [3]. The 5 residents working on the "Foundations to Actions" cohort are currently hosted at the Missouri Botanical Garden, Natural History Museum Los Angeles County, the Ernst Mayr Library of the Harvard University Museum of Comparative Zoology, Smithsonian Libraries and Chicago Botanic Garden.

1.3 Lenhardt Library

The Lenhardt Library is one of the great treasures of the Chicago Botanic Garden. Open to the public 7-days a week, its 125,000 volume collection encompasses resources on gardening, botany, plant conservation and landscape design, in formats from rare books to e-books. The Library supports all aspects of the Garden's mission: "*We cultivate the power of plants to sustain and enrich life*" [4]. The resident hosted at Lenhardt Library (Alicia Esquivel) is dedicated to performing content and collection analyses for BHL.

2 OBJECTIVES

In order to manage the BHL's collection scope, the Collections Committee at BHL created a Collection Development Policy that defines specific areas of interest to digitize based on BHL user needs. The committee defines BHL users as an interdisciplinary audience composed of "zoologists, botanists, evolutionary biologists, taxonomists, systematists, ecologists, natural history collections managers, scientific illustrators, biological science historiographers, and amateur scientists & hobbyists" [5]. Given these user groups, BHL focuses digitization efforts on the core literature of zoological and botanical literature especially those with high concentration of taxonomic names. While BHL can easily calculate the amount of pages and items digitized, it becomes more challenging to calculate the entire scope of biodiversity literature in order to assess which areas are missing in the BHL corpus.

Collections Committee members estimated the core biodiversity literature in 2010 to consist of 495,000,000 pages. This measurement was calculated by estimating the amount of botanical literature in the public domain based on two extensive bibliographies (*Taxonomic Literature: A selective guide to botanical publications and collections with dates, commentaries and types* and *B-P-H: Botanico-Periodicum-Huntianum*). This calculation was then compared to the total amount of botanical species. This ratio of pages to species was then compared to zoological and mycological species counts [6].

This current project explores methods to improve the accuracy of these estimates in order to determine the scope of biodiversity literature both in and out of copyright. Additionally, an aim of this work is to develop visualization methods for examining the coverage within the corpus by comparing BHL materials to known ontologies. This work-in-progress of digital collection analysis explores geographic, taxonomic, and topical tools for analysis. It is performed using BHL metadata such as subject headings and taxonomic name data, as well as exploring full text indexing for potential data mining using natural language processing and machine learning.

To visualize geographic representation in BHL, the full text is mined for geographic names, or toponyms, then assigned latitude and longitude coordinates and placed onto a map.

The taxonomic representation in BHL is done using scientific names mined from collection text using the open source Global Names and Recognition and Discovery (GNRD) architecture, a suite of machine learning and named entity recognition algorithms, to extract scientific names to index and attach page records. The list of scientific names can be exported through BHL's public data tables. In order to determine collection coverage of different biological kingdoms (Plantae, Animalia, Fungi, etc.), the list of scientific names can be sorted by using the Global Names Resolver tool, from the GNRD suite. This tool matches names based on different data sources (taxonomic ontologies such as NCBI, ITIS, and Catalogue of Life). By filtering scientific names into kingdom level taxonomic ranks, we can get a more granular look at species coverage in BHL.

Topical analysis of the collection is being done in collaboration with external researchers. The text is analyzed using a topic model and a controlled vocabulary to extract topics from items based on the full text of the items.

The lessons learned from this research will inform the next iteration of BHL in how to provide meaningful data to its users. By exploring collection analysis methodologies of the BHL corpus, biodiversity knowledge can become better connected across open access platforms and databases.

3 ACKNOWLEDGEMENTS

Funding and support for this National Digital Stewardship Residency comes from the Institute of Museum and Library Services.

REFERENCES

[1] Biodiversity Heritage Library. 2015. Strategic Plan: 2015-2017. Available at http://biodivlib.wikispaces.com/About

[2] Biodiversity Heritage Library. About. Available at http://biodivlib.wikispaces.com/About

[3] Meridith Beck Mink, Samantha DeWitt, Christa Williford, Alice Bishop, 2016, Keepers of our digital future: an assessment of the National Digital Stewardship Residencies, 2013–2016, Council on Library and Information Resources, viewed February 6, 2017.

[4] Chicago Botanic Garden, Lenhardt Library. About. Available at https://www.chicagobotanic.org/library/about.

[5] Biodiversity Heritage Library. Collections Development Policy. Available at http://biodivlib.wikispaces.com/Collection+Development+Policy

[6] Garnett, T. 2010. The Domain of Biodiversity Literature: Estimates of Scanning the Core Texts for the Biodiversity Heritage Library. White Paper prepared for BHL Steering Committee.