On Their Terms

Leveraging Biological Taxonomy for a User-Centric Collection Analysis of a Consortial Digital Library

Jacqueline E. Chapman, Digital Collections Librarian, Smithsonian Libraries

Smithsonian Libraries

BHL Biodiversity Heritage Library

Background

The Biodiversity Heritage Library (BHL)

- Established in 2006
- Collaborative digitization program
- 60+ global partners (natural history, botanical, and zoological libraries)
- Free, open access portal
- 50+ million pages
- Primary audience = scientific taxonomists
- 169+ million instances of scientific names

BHL's Complex FAQs

- What are BHL's major gaps?
- What are BHL's collection strengths?
- How well does BHL serve subject specialists?
- What percentage of biodiversity literature is in BHL?

Considerations

Inconsistent data

- differing international standards
- differing institutional practices
- lack of control over ingested content
- changes over time to certain fields
- lack of authority control

Insufficient data

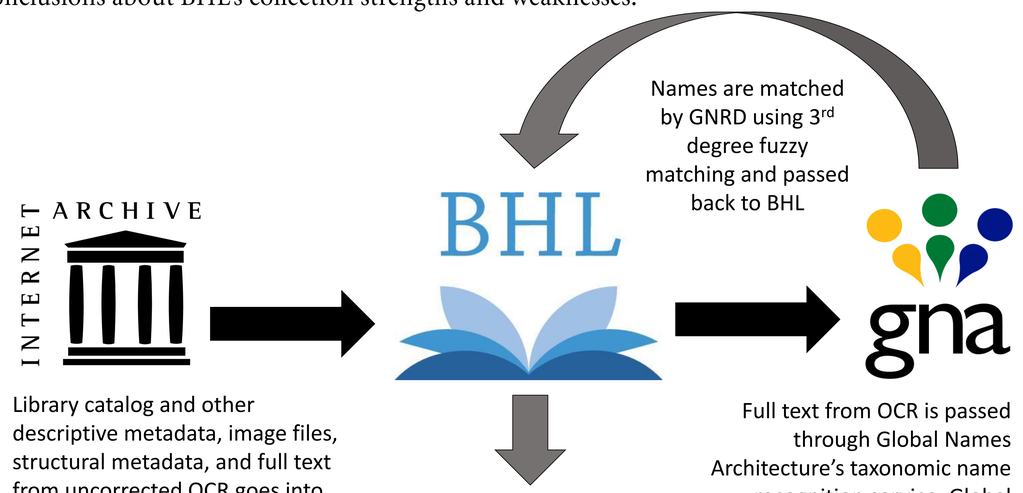
- limited existing item level metadata
- almost no readily available article level metadata
- date (normally year) most often provided at the title level, rather than item level
- subject headings are not granular enough for taxonomists
- subject headings are not specific enough for titles that cover a wide range of topics

Collection Development

- Format (archival material)
- Inadvertent Duplication
- Copyright Status
- Scope (biodiversity literature)
- Completeness (no 'list' to compare against)

The Assessment Possibilities of Full Text

Full-text is a value-add that librarians have long offered alongside the digital surrogates of their physical holdings. BHL has increased the value of its full text derived from Optical Character Recognition (OCR) by matching this text to controlled vocabularies in order to provide users with cross-collection bibliographies for genera and species. This is achieved through a Taxonomic Name Recognition (TNR) service, which provides users with access at the genus and species level. Librarians can use this metadata to assess the taxon distribution in BHL to reach actionable conclusions about BHL's collection strengths and weaknesses.



Analyzing the BHL Corpus Using TNR

To cross-reference the BHL taxonomic names data with an official list of taxonomic names, a list of names must be selected as the control. Unfortunately, GNRD does not have an accessible 'list' of names to draw upon. Part of GNRD's dataset of names comes from the Catalog of Life (CoL), an authoritative global index of taxa with over 1.6 million species described. The resulting data provides an avenue for identifying gaps and showcasing collection strengths, data that is particularly useful when contextualized by subject specialists.

Pilot Study: Ferns and Club Mosses

- Expected to find gaps in the genus coverage, and to use gaps to direct bibliographic analysis
- Instead: very few gaps, and all gaps were accounted for due to copyright limitations
- Out of 306 genus names, only 32 genera appeared fewer than 10 times within the corpus, all of which were described after 1923
- Findings were well received at subject specialist conference

Advantages

Produces metadata that is

- Relevant (taxonomists as primary audience)
- Specific (granular, targeted)
- Consistently applied across the corpus

Utilizes metadata that already exists

- TNR was added to BHL to help users directly, this is a secondary purpose (reuse)
- Scientific Names lend themselves to this kind of work due to consistency across time, geography, and language (Latin names)

Limitations

- Missing names and false positives due to uncorrected OCR, characteristics of the text, and fuzzy matching
- Rapidly changing taxonomic names
- Complex scientific names (e.g. trinomials)
- Not a perfect match between CoL and the names used by GNA's GNRD
- Requires subject specialist knowledge to select the appropriate subset of names and to fully interpret the results

Future Work

Customize Outreach: Provide targeted services and marketing to subject specialists by continuing to assess specific taxon groups, possibly at the species level

Scale Up: Run all of the CoL data over the BHL data to capture a full image of BHL's taxon coverage, possibly at the species level

New Content: Encourage new BHL partnerships with libraries and/or publishers using data to show user needs and scholarly benefits

Quantify Impact: Demonstrate the impact of permissions titles on access to scientific literature by further dividing content