Purpose of BHL Collections Analysis

To determine the extent of the biodiversity literature universe and how much BHL has fulfilled / left to scan?

What are areas of BHL corpus strength and weakness? 1) bibliographic subject matter 2) taxonomic subject matter 3) geographic region 4) language

Bibliographic Subject Analysis

Analyze bibliographic metadata elements for titles:

- What is the overall picture of Library of Congress Subject Headings for the BHL corpus?
- How do the LC subject headings in BHL compare to LC subject headings in the OCLC corpus?
- How do the LC subject headings in BHL compare to LC subject headings for the individual BHL consortium library catalogs?
- Is it possible to compare the subject matter of the BHL corpus against other digital library projects such as Google Books, HathiTrust, and Gallica?
- Can a keyword analysis of BHL title metadata reveal anything about the subject matter?
    - If so, how does the subject matter derived from the title metadata compare to the LCSH headings for the work?
- Maybe compare to JSTOR? What kind of subject metadata do they have?
- Compare LCSH of BHL partner content vs. ingest content

(DPLA? Do they just have links to content or hold actual content?; BTW Europeana looking to expand to becoming a repository of objects not just of metadata[BHLE?])

Taxonomic Subject Analysis

Analyze taxonomic name strings within a given volume:

- What is the overall picture of taxonomic names in BHL? (hundreds of millions of name!)
    - names vs. ”concepts”...
    - Do we have more entomological species than fish species for example? (synonymies could complicate this, number of species existing skews these numbers anyway - extrapolate to a volume count?)
- Is it possible to compare the overall make-up of taxonomic name strings in the BHL corpus to major biodiversity nomenclators such as the Catalogue of Life (http://www.catalogueoflife.org/)?
- Is it possible to determine subject matter based on taxa?
- If possible, then how do the subjects derived from an analysis of the taxa compare to the LCSH assigned to the title?
- Animalbase?
- Compare known species to % represented in BHL - Index Animalium, IPNI, Index Fungorum, WORMS, Avibase, Algaebase, TaxaMatch[1] & IRMNG (see Tony Rees

[1]Rees T (2014) Taxamatch, an Algorithm for Near (‘Fuzzy’) Matching of Scientific Names in Taxonomic Databases. PLoS ONE 9(9): e107510. doi:10.1371/journal.pone.0107510

presentation in Session 03 here: http://gbif.vsworld.com/) ,etc taxon specific databases… nothing all inclusive so repetitive work, need to take respective groups and compare database per database

OCR Text String Analysis

- Can we get a more granular understanding of subjects on a volume-by-volume basis rather than title-by-title? (data exports & APIs have OCR text files)

Bibliography Analyses

- What are the major taxonomic bibliographies that BHL should use to compare its corpus to?
    - UIUC International Field Guide bibliography: http://www.library.illinois.edu/bix/fieldguides/index.html
- How can we compare the BHL corpus to major Botanical and Zoological bibliographies such as Zoological Record, Taxonomic Literature II (TL2), Index Animalium, and the International Plant Name Index (IPNI), botanical journals...(BPH)? Zoobank?
- Journal abbreviation tools helpful here
    - BPH tool: http://fmhibd.library.cmu.edu/fmi/iwp/cgi?-db=BPH_Online&-loadframes
    - JAbbr (Cornell): http://jabbr.mannlib.cornell.edu/about