The Domain of Biodiversity Literature

Estimates of Core Texts for the Biodiversity Heritage Library

Fundraising Issues

May 2010

**Summary:**

BHL has digitized 29,000,000 pages of biodiversity literature. Confining our estimates to botanical, mycological, and zoological literature we have estimated a total of 495,000,000 pages of core biodiversity literature as defined by major compendia such as *TL-2, BPH*, *Index Animalium*, and *Zoological Record*. Our estimates have many assumptions that can be refined for further accuracy as our experience grows. Of the above total, we estimate that 100,000,000 pages are from pre-1923 literature and therefore in principle available for open access scanning. The availability of the post-1923 is hard to estimate because it requires publisher-by-publisher contact and much of it is effectively off-limit for the foreseeable future due to copyright law and commercial interests. Nevertheless, BHL has obtained permissions to digitize backfiles for 90 journal titles and the number is growing. In fact, our ability to scan in a timely manner sometimes cannot keep up with the number of journal titles for which we receive permissions.

From partner projects and other sources the BHL should add approximately 18,000,000 more pages by the end of project year 5. The total cost per page of scanning is ~$.30 per page. Based on the estimates of the core public domain literature, the total cost to digitize the remaining pre-1923 core literature is approximately $10,000,000.

It is estimated that there have been approximately 49,000 not-for-profit or learned society biodiversity journals in existence since 1923. As many as 50% of these may be currently merged or ceased publication.

**Fund Raising Issues:**

1. The BHL has demonstrated to date:
   1. The infrastructure for scanning, mounting, and serving digitized biodiversity literature.
   2. A cooperative working environment involving decision-making, discussion, and work assignment distributed among staff at 12 natural history and botanical libraries.
   3. The provision of web services for reuse of this content by other projects.
   4. A favorable relationship with non-profit scientific publishers and non-profit journal aggregators allowing the BHL to digitize and serve journal backfiles that are post-1923.
   5. A global cooperation to incentivize other national-level projects to host and preserve the content.
   6. The ability to deliver more content through scanning if funding is made available.
2. However, although the BHL currently makes 29,000,000 text pages of biodiversity literature available, this is only 5.8% of the estimated total domain of core biodiversity literature or 29% of the pre-1923 literature.
3. Due to the way in which biodiversity sciences are published, it is very difficult and expensive for the BHL to “target” any particular taxonomic or thematic group for bulk literature scanning in an effective manner. While some journals are focused on taxa, many others are not and will contain mycological, botanical, and zoological articles, e.g. *Journal of the Washington Biological Society.* The existing library metadata is not at a sufficient level of detail to guide fine-tuned selection. Fine-tuning the ability to target is a significant resource allocation. Once digitized and indexed, of course, we can do significant analysis.
4. BHL is adding significant numbers of digitized biodiversity texts in project years 4 and 5 from the following sources:
   1. Ongoing BHL scanning
   2. Already scanned biodiversity texts from BHL members that have not yet been ingested into the BHL Portal
   3. BHL-Europe
   4. BHL-South America
   5. BHL-China
   6. BHL-Australia
   7. Additional ingests from Internet Archive partners with biodiversity literature holdings.
5. These additions to the BHL should bring an additional 18,000,000 pages of pre-1923 core biodiversity texts available through the BHL for a total of 63,000,000 by project year 5 leaving 35,000,000 remaining pages of pre-1923 core biodiversity literature to scan.
6. With the additions in (5.) a donation of ten million dollars will ensure that approximately 100% of the pre-1923 and a significant portion of the post-1923 core literature of biodiversity is made available via open access to the global public using the best standards and accessible to biodiversity projects for reuse. The results of such funding include:
   * 1. A major component of the well-documented “taxonomic impediment” will be eliminated. No longer will researchers have to travel great distances simply to read or copy pages from a book. In addition, this will empower the many competent, worthy taxonomists who are unaffiliated with major research institutions and lack privileges at a research library. This will help accelerate the glacial pace of documenting the world’s biota.
     2. This will hasten the emergence of BHL as the *de facto* starting point for any serious research of the published record of biodiversity. This will create pressures for commercial publishers to want their content aggregated with the content BHL stewards.
     3. Remaining “hold outs” of digitized biodiversity literature, e.g. the Linnaean Society, will be pressured by the market and their boards to share their content with BHL in an open access model. There will be no other alternative.
     4. Developing countries with strong para-taxonomist programs will be able to offer graduate level degree programs since they will have access to a biodiversity library to rival those in existence in the U.S. and U.K. (personal communication from Dr. Bryan Fisher from the California Academy of Sciences who has been running para-taxonomist programs in Peru for many years).
     5. BHL will be able to attract more third party developers for services and data mining.

**Methodology of Estimates:**

1. Estimate the domain of core literature in Botany because it is better documented.
   1. Using *Taxonomic Literature, 2nd Edition* (TL-2), estimate the number of monographs published from 1753-1940 (includes mycology)
   2. Using *Botanico-Periodicum-Huntianum* estimate the number of serials published from 1665 to present. (includes mycology)
   3. Estimate the average number of volumes per title.
   4. Estimate the average numbers of pages per volume based on a sample.
   5. Calculate a total estimated number of pages for the period covered.
2. Estimate the ratio of botanical and mycological species to total scientifically identified species based on *A.D. Chapman’s Numbers of living Species in Australia and the World. 2nd edn (2009).*
3. Using this ratio, estimate the total number of core biodiversity pages
4. Using *Zoological Record* and *Index Animalium* determine the ratio of pre- and post-1923 citations. Use this ratio as way to estimate the number of pre-1923 pages.

**Estimated Page Counts of Taxonomic Literature**

**What we know**

* We can estimate the number of titles of greatest importance to the field of Botany
  + Using Taxonomic Literature, 2nd ed. to estimate the number of monographs published from 1753 to 1940
  + Using the Botanico-Periodicum-Huntianum to estimate the number of serials published from 1665-present
* We can estimate the ratio of Botanical species to total species based on A.D. Chapman’s Numbers of living Species in Australia and the World. 2nd edn (2009).
* We can estimate the average number of volumes per title from our BHL scanning work
* We can estimate the average number of pages per volume based on our BHL scanning work
* We know (discounting the duplicates) the number of pages of biodiversity literature we have scanned to date
* Publishing and learned societies
* Of the 21,000 peer-reviewed journals, monographs and book series listed by Ulrich's Periodicals Directory, at least 9,250 are published by not-for-profit publishers (learned societies, professional associations and university presses). according to the Association of Learned and Professional Society Publishers (ALPSP), the international trade association which represents learned society and other not-for-profit publishers. In addition, many societies outsource the production of their journals to commercial publishers, so it is fair to assume that more than half of all peer-reviewed journals are created by not-for-profit publishers. <http://www.nature.com/nature/focus/accessdebate/8.html>. Using this as a ratio we take the total number of estimated biodiversity serial titles and total number of serial pages to make an estimate for the number of non-profit biodiversity journals and the number of pages they represent.

**Limitations of what we know**

Number of volumes per journal title

Style of publications for botany vs. zoology (serials/monographs)

Existence of core literature outside the sources listed above.

**Costs of Scanning**

The cost per page of scanning includes:

1. direct cost of our scanning partner, Internet Archive for normal pages - .10 per page
2. direct cost of our scanning partner for over size books and fold outs - $1.00 per page
3. For an average of .12 per page, whether normal or oversize from IA.
4. In-house library processing work including selection, conservation review, retrieving from shelves, checking out items, loading and shipping book carts, receiving returned books carts, checking in items, reshelving, and quality control of the digital products. .21 per page.
5. I have rounded this down to $.30 per page because I am assuming some economies of scale for the amounts that are considered here.

However, these estimates have been computed based on the current BHL operations where the following are funded by the MacArthur Foundation or the Moore Foundation and are not part of the $.30 per page estimate.

1. Salaries for the BHL Program Director, the BHL Technical Director, the BHL technical development team, the BHL Collections Manager.
2. Meeting and workshop costs for coordinating a geographically dispersed and technically distributed project involving 12 administratively separate organizations.
3. Administrative expenses for accounting, tracking, reporting, budgeting etc.

Long-term funding for these functions will be addressed, though not solved, in the BHL Sustainability Plan.

**Estimates of time:**

Working primarily with the Internet Archive, BHL can perform the scanning in Washington, D.C., London, New York area, Boston area, and San Francisco. Modest amounts can be worked through the Missouri Botanical Garden also. Assuming funding of $10,000,000 spread over four years and assuming multiple BHL libraries participate an output of over eight million pages per year for four years could be achieved to digitize the remaining pre-1923 biodiversity core literature ~74,000,000. It is not conceivable for the BHL to digitize this amount in less than four years.