Finding scientific articles in a large digital archive: BioStor and the Biodiversity Heritage Library

Roderic D M Page*[1]

[1]Institute of Biodiversity, Animal Health and Comparative Medicine, College of Medical, Veterinary and Life Sciences, Graham Kerr Building, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK

Email: Roderic D M Page*- Roderic.Page@glasgow.ac.uk;

*Corresponding author

Abstract

Background: The Biodiversity Heritage Library (BHL) is a large digital archive of legacy biological literature, comprising over 31 million pages scanned from books, monographs, and journals. During the digitisation process basic metadata about the scanned items is recorded, but not article-level metadata. Given that the article is the standard unit of citation, this makes it difficult to locate cited literature in BHL. Adding the ability to easily find articles in BHL would greatly enhance the value of the archive.

Results: A service was developed to locate articles in BHL based on matching article metadata to BHL metadata using approximate string matching, regular expressions, and string alignment. This article finding service is exposed as a standard OpenURL resolver on the BioStor web site http://biostor.org/openurl/. This resolver can be used on the web, or called by bibliographic tools that support OpenURL.

Conclusions: BioStor provides tools for extracting, annotating, and visualising articles from the Biodiversity Heritage Library. BioStor is available from http://biostor.org/.

Background

In July 2010 Lambert et al. [1] published a paper in *Nature* that described an extinct sperm whale possessing the biggest bite of any tetrapod known. They named this formidable predator *Leviathan melvillei*, the genus name *Leviathan* being derived from the Hebrew 'Livyatan', the species name honouring Herman Melville (author of Moby Dick [2]). As appropriate as this name was, it quickly ran foul of the rules of zoological nomenclature because *Leviathan* had already been used 169 years ago for an extinct species of mammoth [3]. Although the name *Leviathan* Koch [3] had lapsed into obscurity (as a

synonym of *Mammut* Blummenbach) its existence meant the newly discovered whale had to be renamed, which it duly was in a month after the original publication [4].

The fate of Lambert et al.'s *Leviathan* illustrates a significant challenge facing researchers finding and naming new species – the discoverability of existing names. In the absence of a global register of all taxonomic names that have ever been published, a researcher about to publish a new name may struggle to establish that that it hasn't already been used. Zoological nomenclature dates from 1758, botanical nomenclature from 1753, hence a comprehensive list

1

of taxonomic names must survey some 250 years of literature [5], much of which is obscure and may not exist in digital form. Digitising this legacy literature is the goal of the Biodiversity Heritage Library (BHL) [6], a consortium of natural history museum libraries, botanic libraries, and research institutions. The bulk of this digitisation is carried out by the Internet Archive [7], which scans books (broadly defined to include bound issues of journals), creating a set of electronic files for each scanned item, which includes images of individual pages, and text extracted from those pages using Optical Character Recognition (OCR). BHL takes these files (together with the output from the scanning projects of individual BHL members), indexes them by bibliographic metadata and taxonomic names, and makes the content available on its web site [6] (both as web pages and web services). Although the bulk of BHL's scanning activities focus on pre-1923 content that is out of copyright, it has not inconsiderable post-1923 content contributed by its member institutions, notably publications by various natural history museums.

Finding articles in BHL

The BHL archive comprises "items" corresponding to physical objects which are scanned. Items are grouped together into "titles". A single volume book corresponds to a single title and item, whereas a multi-volume work, such as a journal, will comprise several items grouped under the same title (Fig. 1). Noticeably absent from the BHL model is the standard unit of scientific citation, the article. The inability to easily find articles in BHL is a substantial obstacle to integrating this legacy literature into mainstream scientific publishing. The goal of BioStor is to provide tools to locate and extract articles and associated metadata from the BHL archive. To locate articles BioStor makes use of bibliographic databases, assembled by authors, journal editors, and the taxonomic community. The assumption is that metadata for most of the articles of interest will exist in one or other of these databases, hence the task becomes one of matching these bibliographic references to the scanned content in BHL.

For most modern articles the triple of journal name, volume, and starting page is sufficient to uniquely identify an article [8], and tools such as CrossRef's OpenURL resolver [9] can take this this triple and discover whether a Digital Object Identifier (DOI) [10] exists for a that article. Publishers

make use of this tool to map the the literature cited in a manuscript to the corresponding DOI. In an ideal world the BHL model of (title, item, page) (Fig. 1) would map exactly to (journal, volume, page), such that an individual journal would correspond to a title in BHL, and each volume of that journal was a separate item. Given that BHL stores page numbers for each scanned page, locating articles would then be trivial and linking to BHL content could be readily integrated into existing publication processes, as well as bibliographic management tools that make use of CrossRef's services to augment user-provided metadata (e.g., Mendeley [11]).

Unfortunately, the actual mapping between articles and BHL content is often rather more complicated. Large articles (e.g., monographs) may be treated as separate "titles" (effectively as if they were books), rather than parts of the same title. A contributing library may have bound several volumes of a journal together, such that a single "item" may comprise multiple volumes. Volume numbers themselves may not be unique within a journal. The *Annals and Magazine of Natural History* (ISSN 0374-5481), published from 1828 until 1967 (being succeeded by the *Journal of Natural History*, ISSN 0022-2933), is divided into 13 "series", each series numbering its volumes from one onwards. Hence, "volume 1" of *Annals and Magazine of Natural History* may refer to any one of 13 volumes from over a span of 138 years [12]. Journals also differ in whether pagination is unique within a volume, or within parts of a volume. For example, in the journal *Arkiv för Zoologi* (ISSN 0004-2110) each article starts on page 1, so that the triple (*Arkiv för Zoologi*,13, 1) may refer to [13], [14], or any of 23 other articles in volume 13 of that journal.

Discovering articles also assumes that the pagination in BHL is complete and correct, and that one side of a sheet of paper corresponds to a "page". BHL records the page number of regular pages, but not pages that are classified as special in some way, such as title pages, or tables of contents. For example, in page 1 in Lynch et al. [15] is recorded in BHL as being the title page, without any number, which will frustrate efforts to find this article by starting page alone.

While the triple (journal, volume, starting page) is usually sufficient to locate the start of an article, we want to recover all the pages in the article, hence we need both the starting and ending pages. Ideally we could then extract the corresponding set of page

images from BHL and join them together to form an article. However, it is not uncommon for older articles to have discontinuous physical pagination, for example by having plates inserted between pages in the text. In some publications, such as *Isis von Oken*, the text on a page forms two columns, each with its own page number (Fig. **??**), hence one physical page need not equate to a bibliographic page.

Metadata matters

Given that locating articles in a archive of legacy literature such as BHL is a non-trivial task, it is worth considering why such an undertaking is worthwhile, beyond integrating BHL with existing citation practices. Indeed, one could argue that, given that the OCR text for BHL content has been indexed by taxonomic name, the need for indexing by article has been greatly reduced – the user could simply search by taxonomic name and find the content they require. This would be sufficient for many users, especially if we were confident that BHL had correctly indexed all the taxonomic names contained in the pages it has scanned. However, OCR errors mean that a significant fraction of names will be missed [16]. An obvious approach to discovering these missing names would be to take existing databases of taxonomic names and publications and search for those publications in BHL.

Metadata also provides ways for clients to aggregate and filter search results. The Encyclopedia of Life [17] incorporates search results from BHL in its taxon pages, but the user has no obvious means of discovering whether the results are from the same article or not, nor can they order the results by date. As an example of one way the display of search results can be improved by sorting, consider the dispute concerning the correct scientific name for the sperm whale, which is debated in both the scientific literature [18–20] and, more vociferously, Wikipedia [21]. Being able to extract basic metadata from BHL would enable us to visualise the relative popularity of the two alternatives, *Physeter catodon* and *Physeter macrocephalus*, over time (Fig. 3). With the obvious caveat that the literature in BHL is a biased sample of the taxonomic literature, it is clear that *Physeter macrocephalus* is the more commonly used name, but its usage peaked around the start of the twentieth century. By the 1950, the sperm whale was more commonly refer to as *Physeter catodon*. Navigating BHL content by date may help the user

discover why the relative usage frequency of these two names changed in the previous century.

Construction and content

A local copy of the core BHL tables (Fig. 4) was created in MySQL using the data dump provided by BHL(http://www.biodiversitylibrary.org/data/data.zip). Page images and OCR text for individual pages are retrieved as needed using the BHL API and cached locally (together with a thumbnail of the page image).

Locating an article

BioStor provides an OpenURL [22] resolver service to locate articles in BHL. At a minimum the resolver requires the journal name, volume, and starting page of the article being searched for. It may also make use of series and date, if these are provided. This service first checks whether the article already exists in the BioStor database. If the article is not found, the algorithm outlined in Fig. 5 is used to search for the article in BHL.

Step 1 – Finding the journal

The first step is to determine whether BHL includes the journal containing the article. BioStor uses a service provided by bioGUID [23, 24] to find the ISSN [25] for the journal. If the bioGUID service returns an ISSN, the algorithm looks up the ISSN in the `Title Identifier` table (Fig. 4) and retrieves the corresponding BHL `TitleID`. If the bioGUID service doesn't return a ISSN the algorithm attempts to find the journal title to the `ShortTitle` field in the `Title` table using approximate string matching. If it fails to find the title it then searches the `VolumeInfo` field in the `Item` table – for some journals (e.g., *Fieldiana Zoology*, ISSN 0015-0754) the journal title is stored in that field. If at this point we can't find the journal we exit.

Step 2 – Finding scanned items for the journal

Ideally each journal corresponds to a single BHL title, but in some cases the same journal may be represented by more than one BHL title, and hence have more than one `TitleID`. Step 2 uses a hard-coded table of such cases to ensure that all items for a given journal are considered by Step 3.

Step 3 – Finding the volume and page

Ideally the `VolumeInfo` field in the `Item` table would contain just the volume number, however all manner of free-form text may be found there. The volume may be recorded as simple numbers or as strings, sometimes indicating volume, page or date ranges, notes on completeness of the volume, or other comments (e.g., "Index"). Metadata may also be in a variety of languages, such that the field may refer to "Volume", "Band", or "Tome". Nor is metadata always recorded consistently within a journal, for example the `VolumeInfo` field for scanned items belonging to the journal *Proceedings of the Zoological Society of London* contains strings such as:

- Part 1- Part 4 (1833-38)

- 1856

- 1901, v. 1 (Jan.-Apr.)

- Jan-Apr 1906

- 1912 v. 2

- 1923, pt. 1-2 (pp. 1-481)

BioStor uses a set of ad-hoc regular expressions to extract volume (and other information where present, such series, issue, and date) information from the `VolumeInfo` field. If no match to the target volume is found the algorithm exits.

Step 4 – Checking the match

At this stage in the algorithm we will have one or more candidates for the first page in the article. Multiple candidates may occur because the article has been scanned more than one BHL contributor, or because there may be more than one article with the same metadata (see examples of *Annals and Magazine of Natural History* and *Arkiv för Zoologi* discussed above). Some of these matches can be filtered by series or date, if the user has supplied that information. For each remaining match we take the OCR text for the first page in the candidate and compare it to the article title by computing a local alignment between words in the page and word in the title using the Smith-Waterman [26] algorithm. Each pair of words that match exactly are scored $+2$, mismatches, deletions, and insertions are all scored -1. The score for the alignment is normalised by

the match score $\times$ the number of words in the title, so that a perfect match as a score of 1. As an illustration, Figure 6 shows the distribution of alignment scores for the *Annals and Magazine of Natural History*. Most articles in this journal have a score > 0.5, however some articles have very low scores due to poor OCR quality. For example, for "Preliminary notice of the Schizopoda collected by H. M.S. Discovery in the Antarctic region" [27] the corresponding OCR text is "Preltiniiiari/ Xutice of I he Sc/ti:oj/0(/a collcxted hy 11. M.S. ' Dixcovenj' in the Antarctic Rec/io".

Storing articles

Articles extracted from BHL are stored in the same MySQL database that stores the BHL tables, using a simple schema comprising a table for article bibliographic metadata, a table for authors, and a table that joins the authors to the individual articles they've authored. A further table joins the article to the BHL `Page` table (Fig. 7).

Utility and Discussion

The BioStor database is available at http://biostor.org/. It features an OpenURL resolver, and can display individual articles, lists of publications by author, by taxonomic name, and by journal. At the time of writing the database contains 17,858 articles extracted from BHL.

OpenURL resolver

BioStor provides an OpenURL resolver at http://bioguid.info/openurl/. If accessed using a web browser the user is presented with a form where they can enter the bibliographic details of an article, or paste in a full citation and have BioStor attempt to parse it. If the article is already in the BioStor database the article will be displayed, if not BioStor attempts to locate the article in BHL. If it finds potential matches, these are displayed to the user (Fig. 8). For each match the page displays the score based on Smith-Waterman alignment between the page OCR text and the article title. In the example shown in Figure 8, there are three potential matches, two of which have high scores (they are duplicates resulting from two BHL contributors having scanned the same journal). A thumbnail of the first page

in each possible match is shown, the user can click on this to view a larger version of the page if they wish to inspect the match more closely. If they are happy that one of the matches is indeed the article they were looking for they fill in the reCAPTHCA test [28, 29] and click on the corresponding button. BioStor will then retrieve the remaining page images and OCR text from BHL, store the article in its database, then display it to the user.

The OpenURL protocol is supported by a number of bibliographic tools, including EndNote [30] and Zotero [31], so users of those tools can employ the BioStor OpenURL resolver to locate articles. Appending "&format=json" to the OpenURL returns the result in Javascript Object Notation (JSON), hence the service can be used as an API by other developers.

Displaying articles

Articles found by the OpenURL resolver are stored in the BioStor database, and given a unique URL of http://biostor.org/reference/n where n is a unique integer. Figure 9 shows an article [32] being displayed in BioStor. A simple Javascript-based viewer displays a single page as a image, with thumbnails of the all the pages in the article shown in a scrolling list. To minimise the time the article page takes to load the thumbnails are only loaded when visible using a delayed Javascript image loader [33]. The user can navigate through the article by clicking on the thumbnail for a given page. To smooth the transition between individual pages, when the user clicks on the thumbnail for a new page the thumbnail is displayed in place of the full page image while that page image loads. When the page image has loaded the low resolution thumbnail (which will appear fuzzy to the user) is replaced by the higher resolution image, giving the user the sensation that the page has come into focus.

The metadata (such as title, authors, journal name, etc.) can all be edited by the user. These edits will be saved if the user passes a reCAPTHCA test. The metadata can be retrieved in standard formats such as Reference Manager (RIS), Endnote XML, and BibTeX. The web page also contains bibliographic metadata embedded using the Context Object in Span (COinS) technique [34], and `<meta>` tags using the Dublin Core [35] and Google Scholar [36] vocabularies. The article itself can also be downloaded as a PDF file, with bibliographic metadata

embedded using the Adobe's Extensible Metadata Platform (XMP) [37]. Desktop bibliographic software that can read XMP, such as Mendeley [11, 38] and Papers [39], can extract this metadata so that the user need not manually re-enter bibliographic details about the paper.

The article page also displays the taxonomic and, where possible, geographic scope of the article. Taxonomic scope is represented by a tag cloud of the taxonomic names that BHL has found in the OCR text for the article, and by a taxonomic classification of those names based on the 2008 edition of the Catalogue of Life [40]. When an article is added to the BioStor database the OCR text is searched for strings that represent latitude and longitude pairs. Any pairs found are displayed on a Google Map.

Displaying authors

BioStor displays a summary page for each author in the database. To mitigate the problem of an author having more than one spelling of their name, BioStor clusters names using a web service provided by bioGUID [23], which implements Feitelson's [41] weighted clique algorithm for finding equivalent names. The summary page aggregates publications and coauthorships across this set of names. The page uses Exhibit [42] to enable the user to browse the publications by date, journal, and coauthors.

Displaying journals

By default BioStor uses the ISSN to identify journals. Where a ISSN isn't available BioStor uses an OCLC number from the WorldCat service [43]. A user can see all the articles for a given journal by appending the journal's ISSN to the URL http://biostor.org/issn/ (or OCLC to the URL http://biostor.org/oclc/). The resulting web page lists the articles for that journal, as well as a graphical representation of how many articles for that journal have been located in BHL. Figure 10 shows the coverage of the journal *Proceedings of the United States National Museum* (ISSN 0096-3801), published from 1878 to 1968.

Displaying taxonomic names

If the user clicks on a name in the taxonomic tag cloud (Fig. 9), or appends a taxonomic name (or

uBio NameBankID [44]) to the URL http://bioguid. info/name/ for a name that has been taxonomically indexed by BHL, BioStor displays a web page listing the articles in BioStor that contain that name. The page also displays a sparkline showing the distribution of that name over time in the local copy of BHL, and lists taxonomic synonyms of the name according to the Catalogue of Life [40].

Searching and browsing

BioStor supports rudimentary full text search of author names and article titles. It also provides an interactive way to browse articles geographically using Google Maps (http://biostor.org/maps/) (Fig. 11). When the user pans or zooms the map the web page displays the set of articles (up to a limit of 20) that have localities contained within the current bounds of the map.

Future directions

BioStor finds articles by matching existing bibliographies to BHL content, hence it relies on external sources of metadata to locate articles. An alternative approach would be to extract articles directly from the archive. Lu et al. [45] used feature extraction and a mixture of rule-based and machine-learning techniques to extract metadata from BHL OCR text, recovering between 66% to 94% of articles in selection of three journals. The set of articles in BioStor could be used as a training data set to help further develop these methods. Another approach to article extraction is crowd sourcing, where the task of identifying articles would be devolved to users. Ultimately crowd sourcing may prove important in cleaning metadata, but it may prove challenging to engage users in creating metadata from scratch.

The BHL archive has extracted taxonomic names from the OCR text, and BioStor looks for geographic localities encoded as latitude and longitude pairs. We could make more extensive use of the OCR text, for example by extracting citations from the literature cited section of extracted articles using autonomous citation indexing [46]. These citations could in turn be feed into the BioStor OpenURL resolver to attempt to locate them in BHL. The combination of variable citation styles and OCR errors means that the same reference may have be represented by several different citations, requiring tools for cleaning and merging citation data (e.g., [47]).

BioStor is built as a service on the top of a copy of data from BHL, and creates a local bibliographic database of articles. One future direction would be to integrate this data with BHL itself. BHL has an openURL resolver (http://www. biodiversitylibrary.org/openurlhelp.aspx) that primarily supports books rather than articles. Adding metadata from BioStor could enhance the BHL OpenURL service, and provide the biodiversity community with a single source for BHL-derived content. BioStor content could also be added to other bibliographic databases, in particular Mendeley [11, 38]. Mendeley is developing an API for storing and retrieving documents and associated metadata, hence it might be possible to devolve the storing of basic bibliographic metadata to Mendeley, BioStor then becoming simply an OpenURL resolver.

Conclusions

The 31 million scanned pages made available by the Biodiversity Heritage Library (BHL) represents a substantial resource of biological literature. BioStor provides an OpenURL resolver to locate articles in this archive. Each article extracted from BHL is given a unique URL, corresponding to a web page that displays the article pages, and information about the taxonomic names and geographic localities mentioned in the article. BioStor is available at http://biostor.org/.

Availability and requirements

- **Project Name:** BioStor

- **Project Home Page:** http://biostor.org/. Source code is available from http://code. google.com/p/bioguid/source/browse/#svn/ trunk/biostor.

- **Operating System:** The BioStor web site is usable with any modern web browser. The source code can be easily installed on a Mac OS X, Linux server. It has not been tested on a Windows machine.

- **Programming Language:** PHP

- **Other Requirements:** Web server

- **License:** GNU General Public License version 2

- **Any restrictions to use by non-academics:** None

List of abbreviations

If abbreviations are used in the text, either they should be defined in the text where first used, or a list of abbreviations can be provided, which should precede the authors' contributions and acknowledgements.

- API, Application Programming Interface

- BHL, Biodiversity Heritage Library

- DOI, Digital Object Identifier

- ISSN, International Standard Serial Number

- JSON, JavaScript Object Notation

- OCR, Optical Character Recognition

- URL, Uniform Resource Locator

Competing interests

The author(s) declare that they have no competing interests.

Acknowledgements

References

1. Lambert O, Bianucci G, Post K, de Muizon C, Salas-Gismondi R, Urbina M, Reumer J: **The giant bite of a new raptorial sperm whale from the Miocene epoch of Peru**. *Nature* 2010, **466**(7302):105–108.

2. Melville H: *Moby-Dick*. Richard Bentley 1851.

3. Koch AC: *Description of the Missourium, or Missouri Leviathan: together with its supposed habits and Indian traditions concerning the location from whence it was exhumed; also, comparisons of the whale, crocodile and missourium with the leviathan, as described in 41st chapter of the book of Job*. Prentice and Weissinger, 2nd edition 1841, [http://www.biodiversitylibrary.org/item/81522].

4. Lambert O, Bianucci G, Post K, de Muizon C, Salas-Gismondi R, Urbina M, Reumer J: **The giant bite of a new raptorial sperm whale from the Miocene epoch of Peru**. *Nature* 2010, **466**(7310):1134.

5. Anonymous: **The legacy of Linnaeus**. *Nature* 2007, **446**:231–232.

6. **Biodiversity Heritage Library** [http://biodiversitylibrary.org].

7. **Internet Archive** [http://www.archive.org/].

8. Cameron RD: **Scholar-Friendly DOI Suffixes with JACC: Journal Article Citation Convention**. Tech. Rep. CMPT TR 1998-08, School of Computing Science, Simon Fraser University 1998, [http://elib.cs.sfu.ca/USIN/JACC.html].

9. **CrossRef OpenURL** [http://www.crossref.org/openurl].

10. **The Digital Object Identifier System** [http://www.doi.org/].

11. **Mendeley** [http://www.mendeley.com/].

12. Evenhuis NL: **Publication and dating of the journals forming the *Annals and Magazine of Natural History* and the *Journal of Natural History***. *Zootaxa* 2003, **385**:1–68.

13. Alexander CP: **The crane-flies collected by the Swedish expedition (1895-1896) to southern Chile and Tierra del Fuego (Tipulidae, Diptera)**. *Arkiv för Zoologi* 1920, **13**(6):1–32, [http://biostor.org/reference/13820].

14. Michaelsen W: **Neue und wenig bekannte Oligochäten aus skandinavischen Sammlungen**. *Arkiv för Zoologi* 1921, **13**(19):1–25, [http://biostor.org/reference/14784].

15. Lynch JD, Ruíz-Carranza PM, Ardila-Robayo MC: **The identities of the Colombian frogs confused with *Eleutherodactylus latidiscus* (Boulenger) (Amphibia: Anura: Leptodactylidae)**. *Occasional Papers of the Museum of Natural History University of Kansas* 1994, **170**:1–42, [http://biostor.org/reference/228].

16. Wei Q, Heidorn PB, Freeland C: **Name Matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL)**. In *iConference 2010 Proceedings* 2010:284–288, [http://hdl.handle.net/2142/14919].

17. **Encylopedia of Life** [http://www.eol.org/].

18. Holthuis LB: **The Scientific Name of the Sperm Whale**. *Marine Mammal Science* 1987, **3**:87–89.

19. Schevill WE: **Mr. Schevill replies**. *Marine Mammal Science* 1987, **3**:89–90.

20. Schevill WE: **The International Code of Zoological Nomenclature and a paradigm: the name Physeter catodon Linnaeus 1758**. *Marine Mammal Science* 1986, **2**(2):153–157.

21. Page RDM: **Wikipedia as an encyclopaedia of life**. *Organisms Diversity and Evolution* 2010, **10**(4):343–349.

22. de Sompel HV, Beit-Arie O: **Open Linking in the Scholarly Information Environment Using the OpenURL Framework**. *D-Lib Magazine* 2001, **7**(3).

23. Page RDM: **bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics**. *BMC Bioinformatics* 2009, **10**(Suppl 14):S5.

24. **bioGUID** [http://bioguid.info/].

25. **ISSN International Centre** [http://www.issn.org].

26. Smith TF, Waterman MS: **Identification of common molecular subsequences**. *Journal of Molecular Biology* 1981, **147**:195–197.

27. Holt EWL, Tattersall WM: **Preliminary notice of the Schizopoda collected by H. M.S. Discovery in the Antarctic region**. *Ann. Mag. Nat. Hist.* 1906, **17**:1–11, [http://biostor.org/reference/50163].

28. **reCAPTCHA** [http://www.google.com/recaptcha].

29. von Ahn L, Maurer B, McMillen C, Abraham D, Blum M: **reCAPTCHA: Human-Based Character Recognition via Web Security Measures**. *Science* 2008, **321**(5895):1465–1468.

30. **EndNote** [http://www.endnote.com/].

31. **Zotero** [http://www.zotero.org/].

32. Raselimanana AP, Raxworthy CJ, Nussbaum RA: **A revision of the dwarf *Zonosaurus* Boulenger (Reptilia: Squamata: Cordylidae) from Madagascar, including descriptions of three new species**. *Scientific Papers Natural History Museum University of Kansas* 2000, **18**:1–16, [http://biostor.org/reference/50335].

33. **lazierLoad − Javascript Image Lazy Loader for Prototype** [http://www.bram.us/projects/js_bramus/lazierload/].

34. **OpenURL ContextObject in SPAN (COinS)** [http://ocoins.info/].

35. **Dublin Core Metadata Initiative** [http://dublincore.org/].

36. **Google Scholar** [http://scholar.google.com/].

37. **Adobe XMP** [http://www.adobe.com/products/xmp/index.html].

38. Henning V, Reichelt J: **Mendeley - A Last.fm For Research?** 2008:327–328.

39. **Papers** [http://mekentosj.com/papers/].

40. **The Species 2000 and ITIS Catalogue of Life** [http://www.catalogueoflife.org].

41. Feitelson DG: **On identifying name equivalences in digital libraries**. *Information Research* 2004, **9**, [http://informationr.net/ir/9-4/paper192.html].

42. **Exhibit: Publishing Framework for Data-Rich Interactive Web Pages** [http://www.simile-widgets.org/exhibit/].

43. **WorldCat.org: The World's Largest Library Catalog** [http://www.worldcat.org/].

44. **Universal Biological Indexer and Organizer (uBio)** [http://www.ubio.org/].

45. Lu X, Kahle B, Wang JZ, Giles CL: **A metadata generation system for scanned scientific volumes**. *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* 2008, :167–179.

46. Lawrence S, Giles CL, Bollacker K: **Digital libraries and autonomous citation indexing**. *IEEE COMPUTER* 1999, **32**(6):67–71.

47. Councill IG, Li H, Zhuang Z, Debnath S, Bolelli L, Lee WC, Sivasubramaniam A, Giles CL: **Learning metadata from the evidence in an on-line citation matching scheme**. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA: ACM 2006:276–285.

48. Pocock RI: **On the Arachnida taken in the Transvaal and in Nyasaland by Mr W. L. Distant and Dr Percy Rendall**. *Ann. Mag. nat. Hist.* 1898, **1**:308–321, [http://biostor.org/reference/52084].
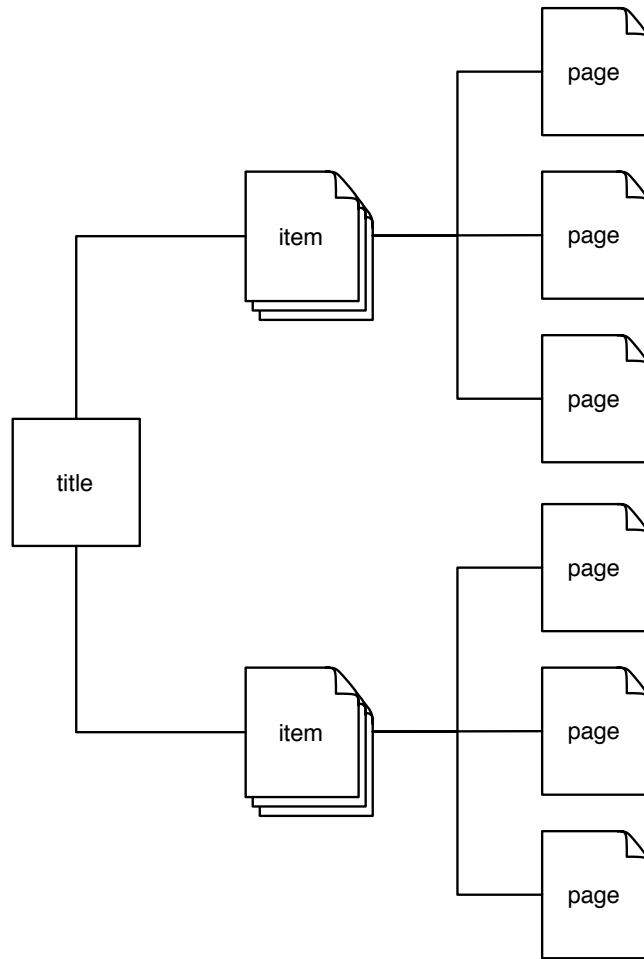
Figure 1:

Figures
Figure 1 - Simplified model of Biodiversity Heritage Library content
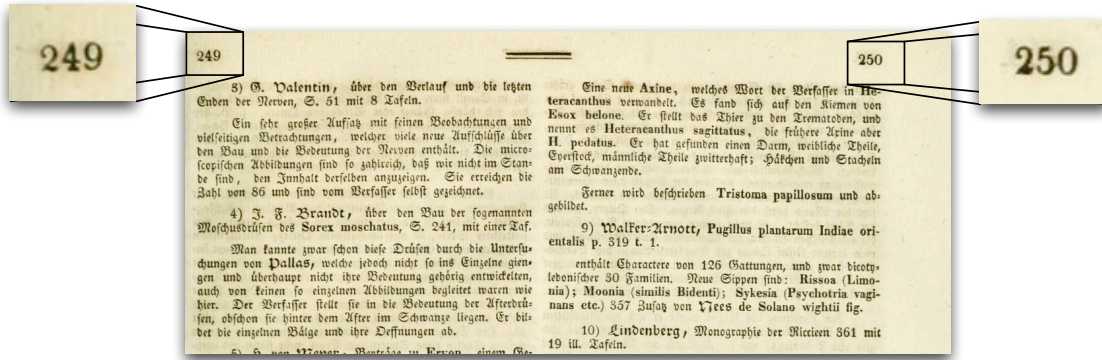Each scanned item comprises one or more page images. Items are grouped together into titles.

Figure 2:

Figure 2 - Physical page with two page numbers

Example of a physical page in the journal *Isis von Oken* with two columns, each of which as its own page number (249 and 250, respectively)
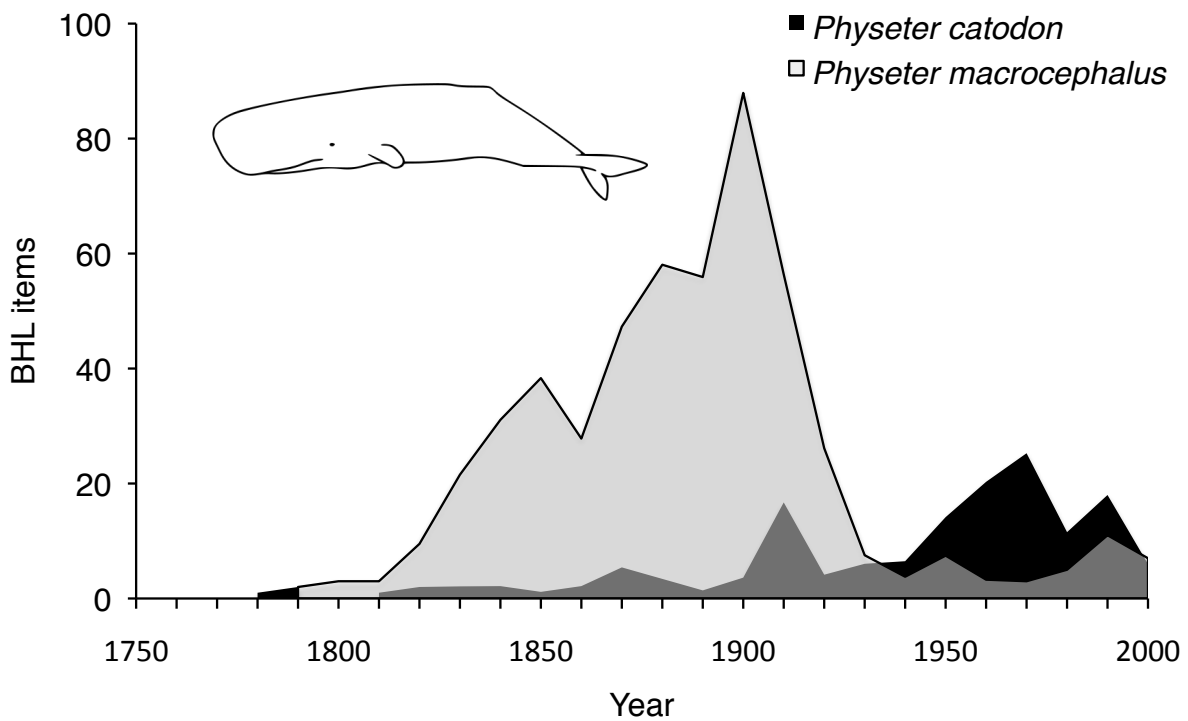
Figure 3:

Figure 3 - Usage of two names for the sperm whale over time

Approximate distribution over time of two alternative names for the sperm whale (*Physeter catodon* and *Physeter macrocephalus*) in items scanned by the Biodiversity Heritage Library. Date of publication was extracted from the StartYear and EndYear fields of the Title table (see Fig. 4).

Page	
PageID	Primary key
ItemID	Item containing this page
PageNumber	Page number
PageTypeName	Page type

Item	
ItemID	Primary key
TitleID	Title containing this item
VolumeInfo	Volume, issue, date, etc.

Title	
TitleID	Primary key
ShortTitle	Short version of title
FullTitle	Title
StartYear	First year published
EndYear	Last year published

Title Identifier	
TitleID	Title
IdentifierName	OCLC, ISSN, etc.
IdentifierValue	Identifier

Figure 4:

Figure 4 - Simplified BHL schema

Simplified database schema for the core tables in the Biodiversity Heritage Library. The fields referred to in the text are shown, together with a brief explanation of their contents.
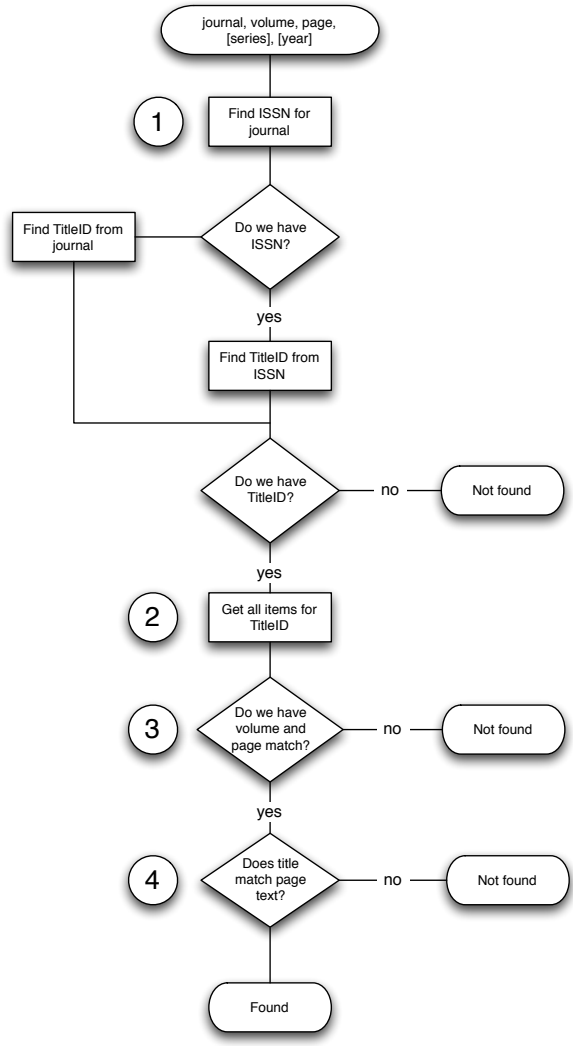
Figure 5:

Figure 5 - Flow chart of algorithm for finding an article in BHL
Steps 1-4 are explained in the text.
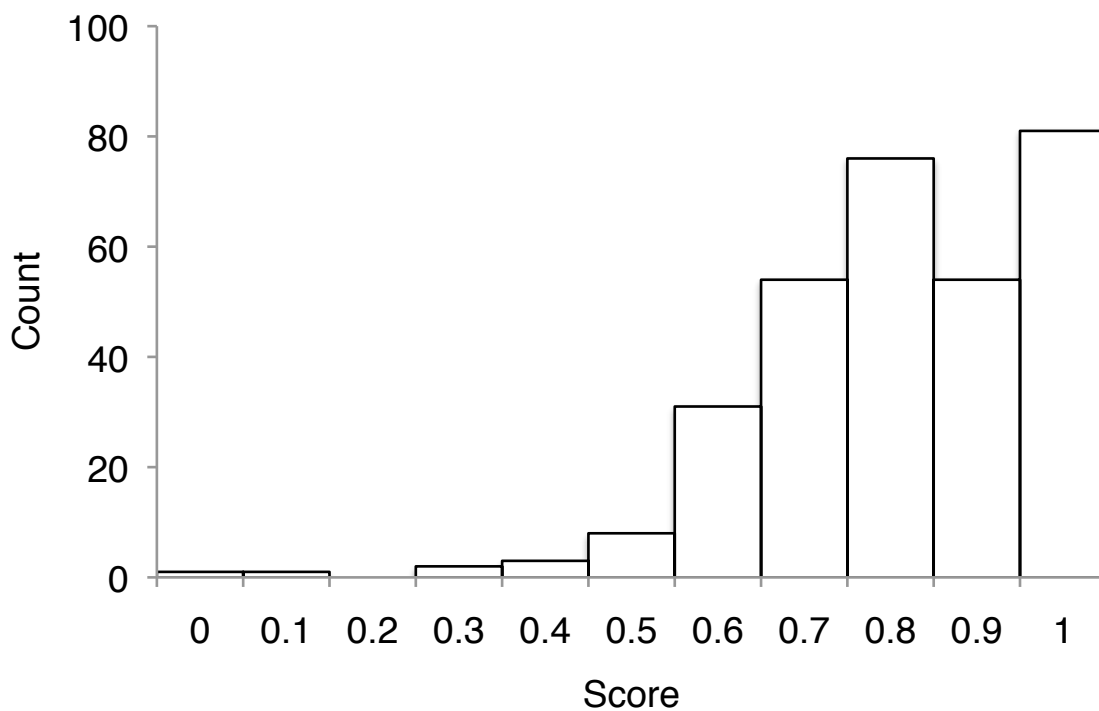
Figure 6:

Figure 6 - **Alignment scores for Annals and Magazine of Natural History**

Frequency distribution of scores for Smith-Waterman alignment between article title and OCR text for 314 articles from *Annals and Magazine of Natural History* in the Biodiversity Heritage Library.

Reference Page Joiner	
reference_id	BioStor reference
PageID	BHL page
PageOrder	Page order

Reference	
reference_id	Primary key

Reference Author Joiner	
reference_id	BioStor reference
author_id	BioStor author
author order	Author order

Author	
author_id	Primary key
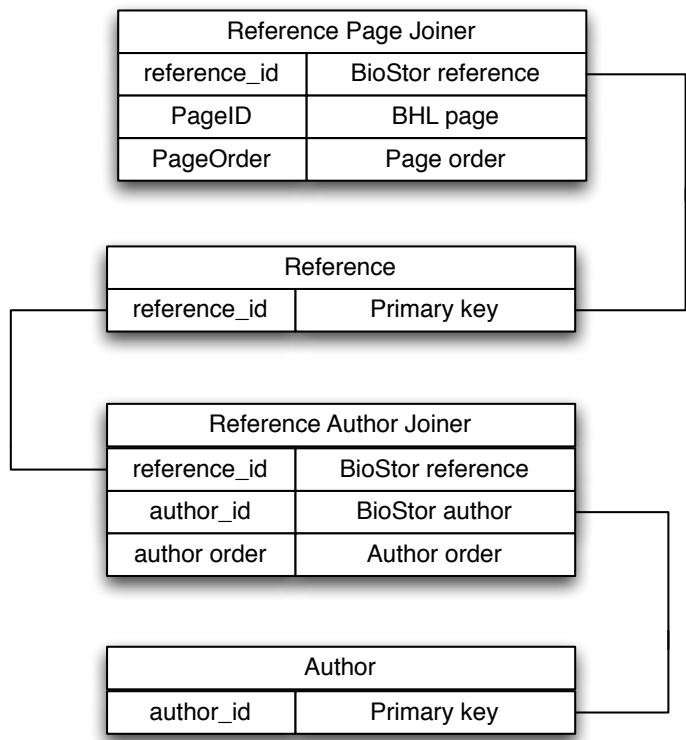
Figure 7:

Figure 7 - **Simplified BioStor database schema**

Reference Finder results

...under ordinary conditions, can be handled with impunity. Lll. — **On the** Arac.hdda laJcen **in the Transvaal and in** ISyasaland hy Air. **W. L. Distant and** I)r. **Percy RendalL** By E,. I. Pocock. Order SCOEPIONES. Family Scorpionidae. OpisthophthaJmus glahrifrons,...
Title match score = 0.52631578947368
BHL PageID 15627941

Click here to accept this match

...ordinary conditions, can be handled with impunity. LI I. — **On the Arachnida taken in the Transvaal and in** I^yasaland ly **Mr. W. L. Distant and Dr. Percy Rendall.** 13y R. I. Pocock . Order Scorpiones. Family Scorpionidae, Opisthophthahnus...
Title match score = 0.84210526315789
BHL PageID 19365519

Click here to accept this match

Mr. Walker **on the** British Chalcidites. 2 U8 et 3 US longiores, 4 US...
Title match score = 0.10526315789474
BHL PageID 2330976

Click here to accept this match

mandiret The

Type the two words:

reCAPTCHA™
stop spam.
read books.

Figure 8:

Figure 8 - BioStor OpenURL resolver

Results of using the OpenURL resolver to find "On the Arachnida taken in the Transvaal and in Nyasaland by Mr W. L. Distant and Dr Percy Rendall" [48] in BioStor. The resolver has found three possible matches in the Biodiversity Heritage Library. For each match the best alignment between the article title and the OCR text is highlighted in yellow.

16

Figure 9:

Figure 9 - Example of page displaying an article in BioStor

The article being displayed is [32].
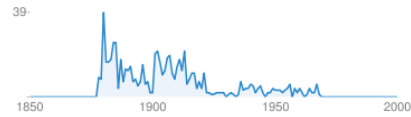
Figure 10:

Figure 10 - Summary of coverage of the journal Proceedings of the United States National Museum in BioStor

Dark blue bars represent pages that have been assigned to an article in BioStor. A sparkline depicts the distribution of these articles over time.
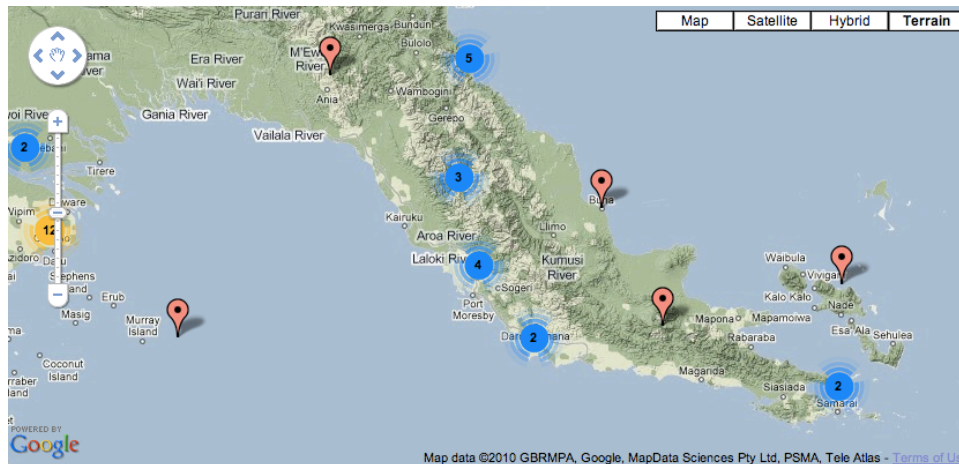
(-11.08902924217906,142.76037597656253)(-6.748722276987813,151.78015136718753)

Reference

1. [A revision of the Pacific species of Conocephalus Thunberg (Orthoptera: Tettigoniidae)](#)
2. [Chelid turtles of the Australasian archipelago: I. A new species of Chelodina from southwestern Papua New Guinea](#)
3. [New Guinean reptiles and amphibians in the Museum of Comparative Zoology and United States National Museum](#)
4. [Grenadiers (Pisces, Gadiformes) from the continental slope of western and northwestern Australia](#)
5. [Bats (Mammalia: Chiroptera) from Indo-Australia](#)
6. [A revision of the "acaecate" earthworms of the Pheretima group (Megascolecidae: Oligochaeta): Archipheretima, Metapheretima, Planapheretima, Pleionogaster, and Polypheretima](#)

Figure 11: .

Figure 11 - Browsing BioStor content geographically using Google Maps
Listed below the map are the articles in the BioStor database with localities contained within the geographic area being displayed in the map.