Finding agriculture among biodiversity: metadata in practice

Jane Bromley, David King, and David R. Morse

Department of Computing and Communications,
The Open University,
Milton Keynes,
MK7 6AA, UK
{j.m.bromley, david.king, david.morse}@open.ac.uk

Abstract. The breadth of biodiversity literature available through the Biodiversity Heritage Library (BHL) is potentially of great use to agricultural research. It provides access to literature drawn from across the world, and its archives document the Earth as it was one hundred years ago and more. However, this strength of BHL is also its weakness: the breadth of coverage of BHL can complicate finding relevant literature. In this short paper, we will explore the practical issues arising from attempting to filter out relevant legacy literature to support agricultural research.

Keywords: agriculture, biodiversity, metadata, AGRIS, AGROVOC, agrotags, KEA, BHL, LCSH, search, keywords, subjects, classification, information retrieval

1 Introduction

The work described in this paper comes from the EU FP7 funded agINFRA project [1], which aims to promote data sharing in agricultural sciences. We are seeking to enhance an existing specialist agricultural resource, AGRIS [2], with content from a more comprehensive – but general – resource, the Biodiversity Heritage Library (BHL) [3], without introducing too many items that are irrelevant to agriculture. In doing this we are not attempting to develop new filtering algorithms. Rather our core task is to create a simple workflow to harvest and filter relevant content from BHL to make it accessible through AGRIS.

We describe how we use AGROVOC [4], a specialist agricultural controlled vocabulary, to assist in accurate filtering of BHL content, and how these vocabulary terms both help and hinder that process. The issues that we are addressing throughout this paper are "what is a suitable list of terms to use to filter?" and "what should we filter on – provided metadata such as the title, classification and subject, or the whole text?"

A brief overview of the relevant repositories and workflows follows.

AGRIS. The UN Food and Agriculture Organization's AGRIS (International Information System for Agricultural science and technology) is a mainstay of agriculture research. AGRIS began in 1976 as a bibliographic reference library to which all interested researchers could contribute, promoting access to agricultural information. It now has more than seven million references, and links to relevant data resources on the web.

AGROVOC. To complement AGRIS, FAO developed AGROVOC, a controlled vocabulary to be "used by researchers, librarians and information managers for indexing, retrieving and organizing data in agricultural information systems and web pages". The consistency provided by using a specific set of defined terms to access agricultural information, including AGRIS, assists productive use of that information. Applying AGROVOC terms to filtered BHL content exposed through AGRIS brings the benefits of discoverability through linked open data to that content.[1]

Biodiversity Heritage Library. The BHL is a large digital archive of legacy biodiversity literature, comprising (in July 2014) over 44 million pages scanned from books, monographs, and journals. The BHL project began in 2005 when ten natural history museum libraries, botanical libraries, and research institutions in the UK and the USA agreed to collaborate in digitizing their legacy literature [5], with texts dating back as far as the c16th. It now draws on libraries "that cooperate to digitize and make accessible the legacy literature of biodiversity" from all of the inhabited continents [6].

Complementing the public domain literature in their collections, the BHL partners have obtained permission from publishers to digitize and publish significant copyrighted content. In conjunction with the partners' geographical scope, this makes the BHL a valuable resource of accessible biodiversity literature. This long-term view can prove invaluable in locating wild relatives of crops and understanding their relationship to local habitats and ecosystems.

Workflow – filtering BHL. BHL's metadata is available as a download [7], updated monthly. Our workflow processes the downloaded metadata to identify agriculturally relevant content, for which we then request the full bibliographic record directly from BHL using its public API [8]. We then pass those records to FAO who imports them into AGRIS. The workflow uses Python scripts that will be freely available on completion of our work.

Related work. Previous writings about BHL's metadata [9] do not discuss its utility. Instead, they consider only the practical problems of assigning metadata to BHL con-

[1] We acknowledge there are related filtering options we could use, eg CAB thesaurus and NALT. However, AGROVOC has the advantage of being an enhancement to AGRIS. In the future, GACS should supersede AGROVOC but is still in development.

tent, given the need to maintain a high throughput in the digitization process, and handling the vagaries of historic biodiversity literature such as separate foldout pages. The current paper does not address these digitization workflow issues.

Previous writings about BHL's content have focused on specific tasks such as named entity recognition [10] within its content in order to improve information retrieval. The current paper does not investigate such methods to enhance retrieval, but to filter BHL content using existing content and search capabilities.

2 Filtering options

This section discusses using four sources of data available to filter out agriculturally relevant resources from BHL, beginning with the item's title, then considering the metadata attributes in subject and classification, and finally using the whole text itself.

2.1 Filtering on Full Title

Our initial approach to filtering was to look for AGROVOC terms in titles because this is the one data source we knew would always be available. Using the October 2012 BHL data export of 56,568 titles, we found that 85% (37,793 titles) of English-language titles contained at least one AGROVOC term, and 73% (41,455 titles) of all titles contained at least one AGROVOC term. This initially promising result masked three problems.

First, 85% (or even 73%) seems a high estimate for the proportion of titles in BHL that are agriculturally relevant. Reviewing the first 20 titles identified suggested they could be appropriate. However, reviewing the top 5 terms that matched[2] to something in a title (birds, plants, history, animals, species) struck us as not particularly *agricultural*, and indeed led to inappropriate titles being selected for inclusion in AGRIS, as identified when we manually reviewed the complete list of filtered titles.

Second, we were not the first researchers to find that AGROVOC can be very broad, and consider that a smaller, focused set of terms could be more discerning. ICRISAT (The International Crops Research Institute for the Semi-Arid Tropics) led the work [11] that produced AGROTAGS [12], a subset of AGROVOC. With some minor edits to aid matching of AGROTAGS terms in item titles, such as removing brackets from the terms, we applied an edited AGROTAGS list as a filter to BHL and retrieved a list 17,670 English-language titles. While the AGROTAGS results list is about half the size of the AGROVOC results list, similar issues in the filtered titles relevance to agriculture emerged when we reviewed the output.

Third, the underlying issue affecting our use of both AGROVOC and AGROTAGS to filter BHL titles is that terms in both lists are not unique to agriculture. An example of the many rather general terms is dry season. In reviewing the accuracy of the results lists we found that a human, reading just the title, cannot tell if the material is

[2] The matching algorithm applies the list of AGROVOC terms alphabetically and stops when a match is found somewhere in the title. Hence, this is not a proper frequency analysis.

relevant: hidden away in the text can be relevant and useful information that is not explicitly explained in the title. Unfortunately, from our perspective, we were not dealing solely with tightly descriptive scientific publications having meaningful titles: BHL's content is broader than that. However, this is a well-known phenomenon, for which the solution is to assign keyword metadata to each item. Therefore, we next investigated the use of keywords to filter BHL content.

2.2 Filtering on Subject

Keywords, or subjects as BHL names them, aid searching for relevant material. Works can have multiple subjects assigned to them and are intended to indicate the content of a work. Hence, they could be used for filtering, as well as searching, BHL.

The subjects applied to BHL's content come from several sources. Most subjects are already associated with the material in the donating institution and are added to the content's BHL metadata as part of the digitization process. These subjects typically include the Library of Congress Subject Heading (LCSH) [13], because many consortium members are libraries and curate their collections using this system. Additional ad hoc metadata can be supplied manually during the digitization process and after. Text mining is used to identify taxonomic names automatically in the content, providing another means to search the literature [14]. Other potential search criteria, such as people and places, are not currently automatically identified using text mining, though BHL would like to enhance its workflow to include this process. The net result is that currently, the metadata subjects might not provide a comprehensive insight into the content of a document.

The July 2014 BHL export data shows that 72,034 out of 77,552 (or 92.88%) titles have subject words associated with them.. That this value is slightly down on typical metadata completeness, as reported for example by Tsiflidou and Manouselis when assessing metadata tools [15], is a product of the BHL data import process. However, we felt there is sufficient coverage of the content to make selection on subjects a valid filtering technique.

Our filtering is based on looking for appropriate LCSH terms in the BHL Subjects field. This is effective not only because the majority of BHL's contributing libraries manage their collections using the Library of Congress cataloging scheme, but where other schemes are used, the subject terms are broadly similar. An added benefit of using LCSH to filter the titles is that the Subjects associated with each title can be translated to AGROVOC using the mapping developed as part of the AGROVOC Linked Open Data project [16, 17]. This mapping simplifies the integration of BHL metadata into AGRIS.

Using monocot as our topic of interest identified seven BHL titles. Interestingly, four titles used the LCSH preferred term 'Monocotyledons', while three used the older and still recognized though no longer to be used 'Monocotyledones'. This offers the possibility of automating record curation before import into AGRIS, bringing all seven items in line with current usage, though at the expense of maintaining detailed matching lists of variant terms. Therefore, we considered exploiting the hierarchical

nature of LCSH, and to select titles based on higher-level subjects only. In this example, the broader term for Monocotyledons is Angiosperms.

We experimented by just using the high-level LCSH term “Agriculture” to filter the Subject field, which returned 2,123 titles (2.74%) as relevant. Filtering the Subject field using the wider criteria of any term that includes the word “agriculture” returned 2,314 items (2.98%), while using “agricultural” returned 881 items (1.14%).

Repeating the experiment starting with a narrow filter using just the LCSH term “Horticulture” returned 3,047 titles (3.93%) as relevant. Filtering on the wider criteria of terms that include the word “horticulture” returned 3,834 items (4.94%), while using “horticultural” returned 83 items (0.11%).

There was some overlap in these results, hence the total titles filtered as agriculturally relevant using LCSH “Agriculture” and “Horticulture” was only just over 4,000 titles. This represents less than 10% of the content of BHL, which was surprisingly small given the large number of titles previously identified as potentially agriculturally relevant.

Many titles have very restricted Subjects, e.g. Banana is used as the Subject for five items and none of them would have been retrieved if “Agriculture” or “Horticulture” had been used as the filter term because they did not have these higher-level terms in their Subject fields. The issue can lead to relevant titles being hard to select. While using a few common LCSH terms, such as “Agriculture”, seems a good way to filter it means that many useful items are missed. Therefore, we next considered returning to a larger list of LCSH terms, but only agriculturally relevant ones.

To achieve this goal we exploited the set of LCSH terms that map directly to an AGROVOC term. Reviewing the list of mapped terms did not induce confidence because it contains generic terms such as “Bread”, “Density” and “Mouth”. Therefore, we began to contemplate repeating the manual curation adopted by ICRISAT when producing AGROTAGS, and to produce our own list of terms to filter BHL Subjects.

We began to prepare a hand-crafted list by reading LCSH and selecting all related terms in suitable hierarchies, but soon realized this would not be sustainable beyond the end of the agINFRA project to accommodate updates to LCSH. In addition, our discovery work with BHL content quickly exposed items whose Subjects do not reveal the true content of the item. For example, David Livingstone’s *Missionary travels and researches in South Africa* [18], has the following subjects: 1813-1873; Description and travel; Livingstone, David; Missions; South Africa; Travel. Neither the title nor the Subject list suggest that this is relevant to agriculture, but the table of contents shows: domestic animals, The Boers as Farmers, Discovery of grape-bearing vines, The sugar-cane, Coffee Estate, Coffee Plantations amongst others.

Therefore, we turned to another means of identifying books used by librarians, and which should also benefit from their curation of the titles before submission to BHL.

2.3 Filtering on Classification

Libraries assign a unique Classification to the items in their collection using a Classification Authority, such as Library of Congress Classification LCC and Dewey Deci-

mal Classification DCC. For example, LCC classifications starting with S mean the item is Agricultural, as does DCC 630. This Classification also informs the item's shelf-mark or Call Number, which is the physical location of the item. Being unique for each item, we investigated the utility of the Classification for filtering material.

In the July 2014 BHL data export, around 13,000 items have LCC Agriculture. Hence, selecting items with LCC Classification "Agriculture" would net just over 15% of the content of BHL. Unfortunately, the classification is not always present. In the July 2014 data export, only 43,848 of the titles in BHL (56.54%) have a Classification associated with them. Therefore, this could be good enough to filter BHL material as a rough and ready method with a high proportion of the retrieved titles being relevant, but would fail to retrieve many relevant titles. In technical terms, filtering on Classification would easily deliver high precision but with low recall [19].

Further, we have found we would be at the whim of each library's local practice. It is up to each library to assign a Call Number, which may include looking at the title, introduction and content, while using a classification manual for guidance. The final choice is up to the particular library; and depends on things like its particular size and remit. So, for instance, what one library places on the Agriculture shelf, another may place under Horticulture, or Economics. Hence, *L'Illustration horticole* [20] is listed under Botany (LCC QK), though it contains useful information about Floriculture, Gardening, Greenhouses and Horticulture. Should we filter on QK, however, we would identify many items that are not agriculturally relevant because the Classification 'Botany' is too broad a term.

Returning to our Subject example in the previous section, of the five texts with the Subject Banana, four are classified as S, Agriculture, and one as QH, Natural history - Biology. Hence, filtering purely with the Classification S can produce relevant results, but it is not sufficient to identify all relevant literature. Yet filtering including the more general QH will prompt the retrieval of a wide range of natural history material not relevant to agriculture.

Therefore, it appears that filtering on terms in titles and subjects, and by classifications lead to the same problem: with a narrow set of terms we can achieve high precision but poor recall, and as we widen the set of terms used so precision falls off to such a degree as to invalidate our filtering. Hence, we turned to another source of information, the full content itself.

2.4 Filtering on the whole text

Given the accuracy and completeness issues with using metadata to filter relevant literature, we have begun to explore the option of filtering based on analyzing the whole text. This approach has two disadvantages. Firstly, we need access to the whole text not just the metadata export. Following an earlier collaboration with BHL, we have a 5Gb local copy of sample articles for our research. This avoids the issue of downloading and analyzing the text of BHL content during our research, though the issue of access remains for any possible later harvesting of the full BHL content. Secondly, there is the issue of the processing power required by the filtering process when compared with the previous metadata-based approaches.

We analyzed the whole text of a sample of articles using KEA [21], a keyword extraction tool trained to apply the AGROVOC vocabulary to the analyzed text. Applying this approach to analyzing the agriculturally relevant book, *The arthropod fauna of potato fields* [22], KEA identifies the key subjects as Arthropoda; Agriculture; Canada; Control methods; New Brunswick; Research; Species; Fields; North America; Yields. This indicates it is a relevant item. In contrast, the BHL supplied key subjects Arthropoda; New Brunswick; Nomenclature, do not indicate the book's agricultural relevance. This suggests that whole text processing can be useful so we are continuing to develop this promising approach to filtering.

A probable refinement to our workflow will be to continue with a first level analysis of the metadata using a narrow set of manually curated AGROVOC terms to identify immediately relevant literature, and only incur the overhead of accessing and processing the whole text for the remaining texts whose relevance we cannot confidently determine solely from metadata.

3 Conclusion

This paper has documented briefly our experience with filtering the known content of one resource to make a relevant subset available to a new audience, enhanced with its discoverability through linked open data. We faced two key issues in our work.

The first issue was to identify "what is a suitable list of terms to use to filter?" While we are fortunate that our target domain of agricultural research has an established vocabulary, AGROVOC, in practice AGROVOC is too generic to be usefully applied unmodified to BHL. The degree of modification necessary to achieve high precision with acceptable recall is still under test.

The second issue was to ask "what should we filter on – provided metadata such as title, classification and subject, or the whole text?" The answer seems to be a combination of these data sources for we can relatively quickly assess the relevance of much material solely from its supplied Subject and Classification metadata, leaving a candidate body of material suggested by the presence of AGROVOC terms in the title to be further refined through whole text analysis.

There is an interplay between these two issues, for it is possible that we can accept poor recall when addressing issue one because it can be overcome by our whole text work to address issue two. Our work to develop this complete workflow, and provide metrics on its efficacy, continues.

4 Acknowledgements

The work is supported by agINFRA, a project funded by the EU Seventh Framework Programme (FP7) under objective Infra 2011 1.2.2, Data infrastructures for e-Science. Grant agreement no 283770. We are grateful to Valeria Pesce and Fabrizio Celli for their help in defining subsets of AGROVOC terms and importing references into AGRIS, and Guntram Geser for discussions.

References

1. agINFRA, http://aginfra.eu/, accessed 18 July 2014
2. AGRIS, http://agris.fao.org/es/content/about, accessed 18 July 2014
3. BHL–Portal, http://www.biodiversitylibrary.org/, accessed 18 July 2014
4. AGROVOC, http://aims.fao.org/website/AGROVOC-Thesaurus/sub, accessed 18 July 2014
5. Gwinn, N.E., Rinaldo, C.: The Biodiversity Heritage Library: sharing biodiversity literature with the world. IFLA Journal. 35(1), 25–34 (2009)
6. BHL–Africa, http://blog.biodiversitylibrary.org/2013/04/making-bhl-africa-reality-bhl-africa.html, accessed 18 July 2014
7. BHL–Export, http://biodivlib.wikispaces.com/Data+Exports, accessed 18 July 2014
8. BHL–API, http://biodivlib.wikispaces.com/Developer+Tools+and+API, accessed 18 July 2014
9. Pilsk, S.C., Person, M.A., Deveer, J.M., Furfey, J.F., Kalfatovic, M.R.: The Biodiversity Heritage Library: Advancing Metadata Practices in a Collaborative Digital Library. Journal of Library Metadata. 10(2-3), 136–155 (2010)
10. Wei, Q., Heidorn, P.B., Freeland, C.: Name Matters: Taxonomic Name Recognition (TNR) in Biodiversity Heritage Library (BHL). Paper, iConference 2010 (2010) http://hdl.handle.net/2142/14919
11. Balaji, V., Bhatia, M.B., Kumar, R., Neelam, L.K., Panja, S., Prabhakar, T.V., Samaddar, R., Soogareddy, B., Sylvester, G.A., Yadav, V.: Agrotags – A Tagging Scheme for Agricultural Digital Objects. In: Sánchez-Alonso, S., Athanasiadis, I. (eds.) 4th Metadata and Semantic Research Conference. CCIS, vol. 108, pp. 36–45. Springer, Heidelberg (2010) http://dx.doi.org/10.1007/978-3-642-16552-8_4
12. AGROTAGS, http://agropedia.iitk.ac.in/content/agrotags, accessed 18 July 2014
13. Library of Congress Subject Heading files, http://www.loc.gov/aba/publications/FreeLCSH/freelcsh.html, accessed 18 July 2014
14. BHL–scientific names, http://biodivlib.wikispaces.com/Developer+Tools+and+API#Developer%20Tools-Scientific%20Names, accessed 18 July 2014
15. Tsiflidou, E., Manouselis, N.: Tools and Techniques for Assessing Metadata Quality. In: Garoufallou, E., Greenberg, J. (eds.) 7th Metadata and Semantics Research Conference. CCIS, vol. 390, pp. 99–110. Springer, Heidelberg (2013) http://dx.doi.org/10.1007/978-3-319-03437-9_11
16. AGROVOC Linked Open Data project, http://datahub.io/dataset/agrovoc-skos, accessed 18 July 2014
17. Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbahndari, S., Jacques, Y., Keizer, J.: The AGROVOC Linked Dataset. Semantic Web. 4(3), 341–348 (2013) http://eprints.rclis.org/20648/, accessed 18 July 2014
18. Download from http://www.biodiversitylibrary.org/bibliography/60038#/summary
19. Van Rijsbergen, C.J.: Information retrieval, Butterworths, London (1979)
20. Download from http://www.biodiversitylibrary.org/bibliography/131#/details
21. KEA, http://www.nzdl.org/Kea/, accessed 18 July 2014
22. Download from http://www.biodiversitylibrary.org/bibliography/63088#/summary