Subjectifying Library Users to the Macroscope Using Automatic Classification Matching

Paul Matthew Gooding (paul.gooding@glasgow.ac.uk), University of Glasgow, United Kingdom

Melissa Terras (m.terras@ed.ac.uk), University of Edinburgh, United Kingdom

Linda Berube (l.berube@uea.ac.uk), University of East Anglia, United Kingdom

Mike Bennett (mike.bennett@ed.ac.uk), University of Edinburgh, United Kingdom

Richard Hadden (richard.hadden@ed.ac.uk), University of Edinburgh, United Kingdom

Introduction

Libraries are sources of large-scale data: both in terms of their collections and the information they collate on their spaces, users, and systems. These data provide opportunities to explore technical, methodological, and ethical questions from the valuable interdisciplinary perspective of Data Science and the Digital Humanities. In light of this, we will explore our analysis of library datasets using *Subjectify*[1], an automatic classification matching tool developed to assist analysis of UK Non-Print Legal Deposit (NPLD) collections. NPLD regulations were introduced to the UK in 2013 to support legal deposit libraries to collect electronic publications (2013). Access restrictions mean that readers may only use these materials on fixed terminals within the physical walls of the six legal deposit libraries (see British Library, 2014 for details). The resultant web logs are therefore unambiguous sources of NPLD collection usage within UK legal deposit libraries.

Our study is part of an established tradition of user studies in the digital humanities. To date, these have focused on user behaviour with digital resources (Warwick et al., 2008; Ross and Terras, 2011; Sinn and Soares, 2014). Web log analysis has been used successfully in this context for over twenty years (Almind and Ingwersen, 1997; Nicholas et al., 2005; Gooding, 2016). These studies adopt methodological approaches and topics of study that contribute directly to our understanding of information sources in the digital humanities. However, there have been fewer studies that address critical humanistic perspectives to inform approaches to the data itself. This paper addresses that gap by describing our research into the users of NPLD materials in the United Kingdom, and the implications of automatic classification matching for library dataset analysis. It will address the following questions: what insights into users of digital library collections can be derived from automatic classification matching? What limitations are introduced by the use of existing classification schemes? And, in light of ongoing debates on responsible data curation in DH (Weingart, 2014; Brown et al., 2016), how might DH and LIS scholars collaborate to inform ethical analysis of large-scale library datasets?

Methodology

Our analysis follows Bates' observation that scholarly communication practices function differently across domains, and that "these many differences *do* make a difference" (1998: 1200) to information access and use: as such, we should be able to identify differences in behaviour by studying the subjects requested by users. To this end we were provided with two datasets comprising title-level NPLD access logs from the reading rooms of the six UK legal deposit libraries[2]. The anonymous logs contained only bibliographic records of NPLD materials accessed by users, excluding both identifiable information about users and interactions with

[1] The code and documentation for *Subjectify* is available on Github at https://github.com/mbennett-uoe/librarytools.

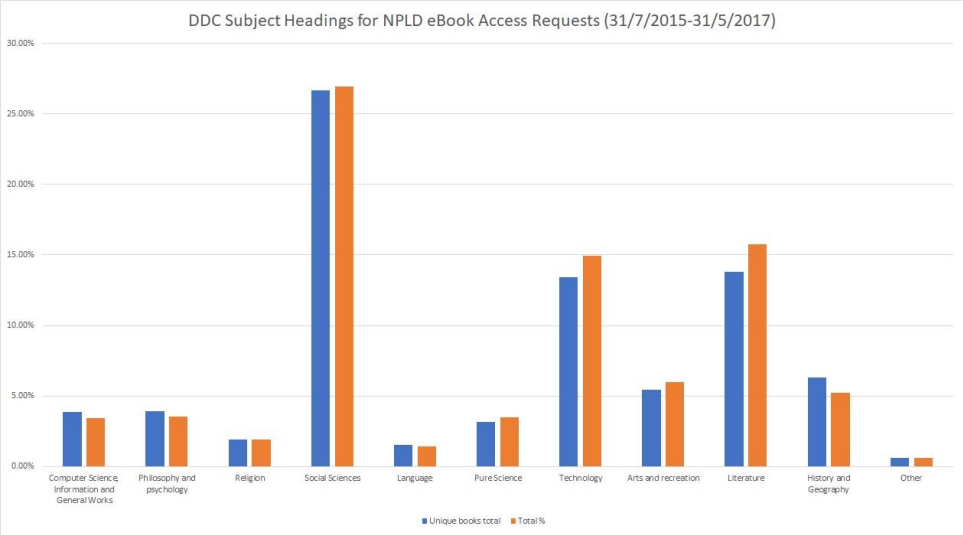
[2] The six libraries are the British Library, National Library of Scotland, National Library of Wales, Bodleian Libraries, Cambridge University Library, and Trinity College Dublin Library.

discovery systems and other materials. The first dataset comprised metadata for all eBook requests (total 91,809) from 31st July 2015 to 31st March 2017. The second dataset comprised metadata for all eJournal article requests (total 36,506) over the same period. Each dataset contained the following metadata: date and time of access request; originating legal deposit library; title of book or article; journal title (where applicable); publisher; and ISBN or ISSN. Each dataset was provided as a CSV file, then cleaned by the research team in OpenRefine to address metadata errors.

Although our dataset contained no identifiable information about users, it may still be possible to infer information about users from the works they consult. We therefore decided to abstract our data and undertake a macroanalysis of user behaviour. To achieve this, we created a small Python-based tool called *Subjectify*. This tool uses the OCLC Classify2 API service[3] to automatically obtain Dewey Decimal (DDC) and Library of Congress (LCC) classmarks from CSV files using key data fields such as title, author, and ISBN. It additionally provides for different options to locate relevant fields to allow input from different data sources. *Subjectify* found a matching classmark for 76.42% for eBooks, and 55.53% for eJournals. This was partly due to missing key data fields in records for eJournals, and partly because many records did not have a corresponding classmark: time-consuming manual classification samples via OCLC Classify2 achieved only slightly higher accuracy rates. We discarded unclassified records used the remaining records to represent patterns of usage by DDC subject. Due to the large number of repeat requests due to system timeouts, we split the remaining records into unique titles (each title counted once regardless of number of requests), and total results (including repeat requests). Our results show that findings were not unduly influenced by repeat requests.

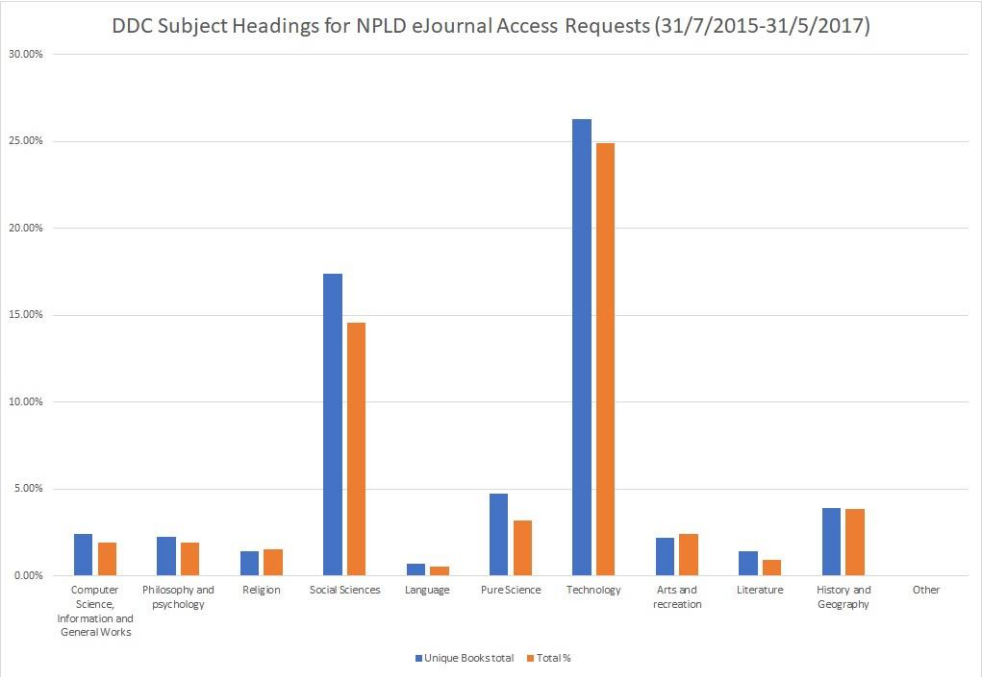
Findings

The following charts show the most commonly accessed subject, by DDC, for titles viewed by users of eBook and eJournal materials. We found that usage by DDC differs distinctly from the spread of classmarks across, for instance, the BL's entire collections[4]:



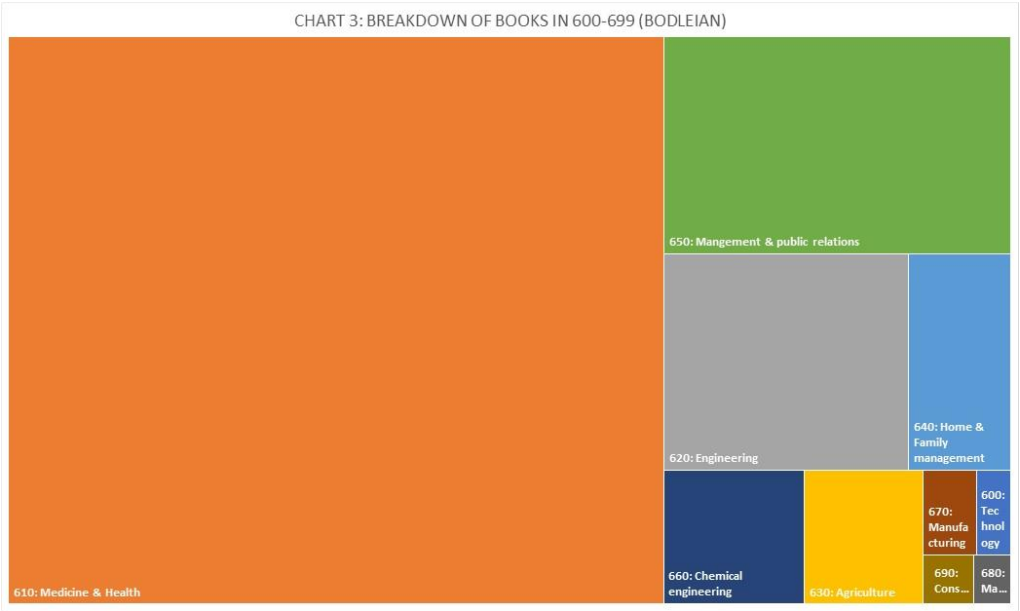
[3] http://classify.oclc.org/classify2/api_docs/index.html

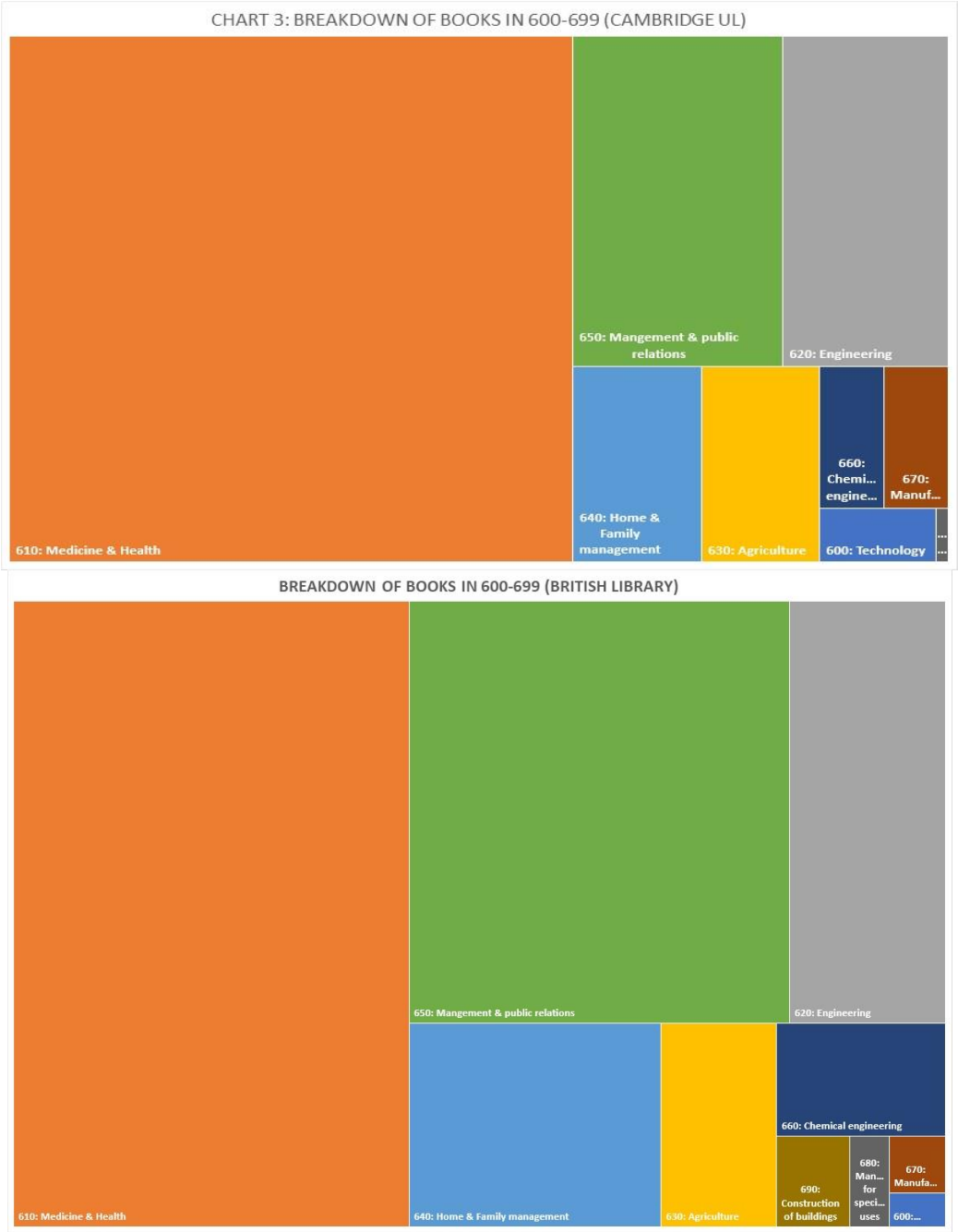
[4] The sub-subroutine blog has produced a fascinating visualisation of the BL's collections, which is worthy of comparison (sub-subroutine, 2015).



Social Sciences texts were notably among the most commonly accessed titles for both formats. The most common subject for eJournals was Technology, whereas for eBooks both Social Sciences and Literature subjects were more frequently accessed. Our findings reflect differing information behaviour across domains: books, for instance, remain a vital source for the Arts and Humanities (Stone, 1982; Palmer and Cragin, 2008) whereas technology and science subjects rely on journals (Talja and Maula, 2003), which tend to provide faster publication of new research. Indeed, Stone's flagship early study noted that "retrospective coverage may be more important to the humanists than having access to current material" (1982: 296). The Social Sciences, on the other hand, are shown by our findings to be more hybrid in their reading patterns.

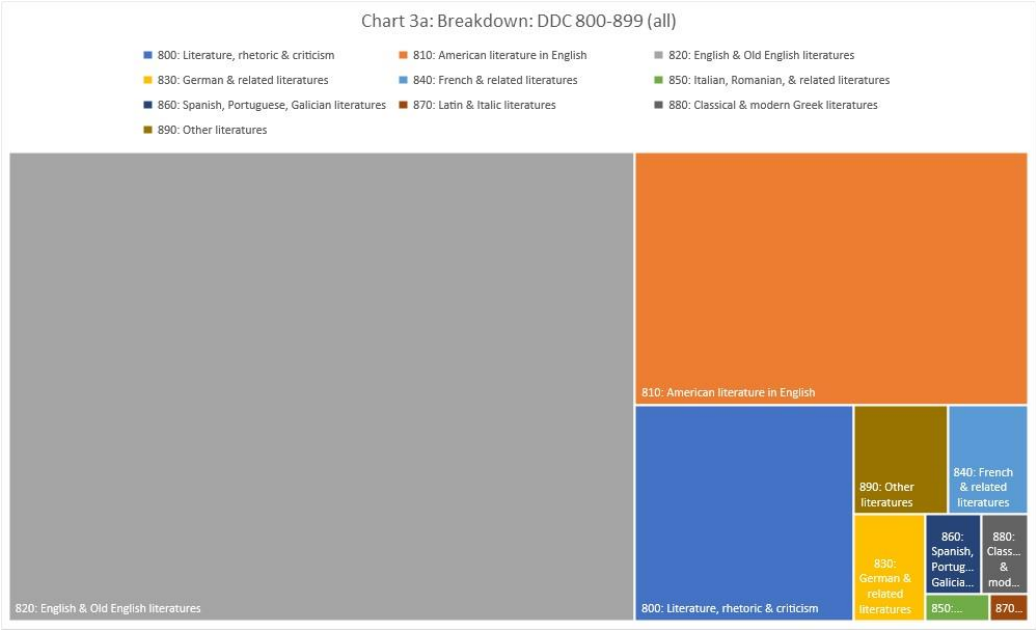
Analysing individual subject categories can derive a better sense of the difference between institutions. The following charts show usage of books in the DDC600-699 classmark (Technology), for the Bodleian Libraries, British Library and Cambridge University Library.





The British Library receives proportionally fewer requests for NPLD medical and health materials. Our interviews with staff at the Bodleian and Cambridge University libraries showed that local medical school staff were a key user group, so we can trace a direct correlation between the user communities of these libraries, and the subjects of the NPLD material used.

Finally, the following table demonstrates usage of 800-899 (Literature) sources to underline a key problem with DDC for automatic classification:



Library classification is a subjective process undertaken by humans within the biased frameworks provided by existing classification schemes (Mai, 2010). Here, for instance, DDC provides distinct categories for English, American, and classical European schools of literature, while lumping the rest of the world under “other literatures” (800-899). This bias emerges from the nineteenth century North American perspective embedded within DDC (Kua, 2008). By relying on automatic matching, we inevitably embed problematic perspectives into our data: while our case study uses UK legal deposit collections, which comprise works represented strongly by DDC, the applicability of this method for library datasets in other parts of the world is questionable. Each time we zoom in with the macroscope, the bias of our chosen classification scheme becomes increasingly evident in the resultant data structures – yet in order to report on literature from without the established canon, we have to do precisely this. The use of established classification schemes is therefore both a methodological and epistemological problem, and future work will be needed to refine our approach.

Conclusion

Our results demonstrate that *Subjectify* was successful in allowing us to analyse user behaviour at scale. It contributes to macroscopic analysis of library data in two ways: first, it allows us to report on bibliographic library users while maintaining privacy through data abstraction; and second, this abstraction allows us to identify patterns of user behaviour with NPLD materials. We believe this approach would work for subject-based analysis of similar collections of bibliographic data, and that it does so in a way that closely reflects how collections are represented in libraries. However, we are also aware that automated classification introduces the biases of those classification schemes into our own data (Adler, 2017). This is an unfortunate side effect of the growing scale of library data. Weingart (2014) notes that the role of the humanities is to tie the very distant to the very close, in order to become ethical stewards of our data. It is therefore essential, when viewing library datasets from a humanistic perspective, to consider the ethics of data representation in our own work. Our priority for future work is to consider how a fruitful conversation between DH and Information Science might develop more nuanced approaches to representing (in the sense meant by Unsworth, 2000) of library data. This should include further consideration of the consequences of how bias in library classification schemes affects microanalytic approaches to bibliographic datasets.

References

Adler, M. (2017). *Cruising the Library: Perversities in the Organization of Knowledge*. New York: Fordham University Press.

Almind, T. C. and Ingwersen, P. (1997). Infometric Analyses on the World Wide Web: Methodological Approaches to 'Webometrics'. *Journal of Documentation*, **53**(4).

Bates, M. (1998). Indexing and access for digital libraries and the Internet: human, database, and domain factors. *Journal of the American Society for Information Science*, **49**(13) doi:10.1002/(SICI)1097-4571(1998110)49:13<1185::AID-ASI6>3.3.CO;2-M (accessed 23 March 2011).

British Library (2014). Legal Deposit *British Library - About Us* http://www.bl.uk/aboutus/legaldeposit/.

Brown, S., Clement, T., Mandell, L., Verhoeven, D. and Wernimont, J. (2016). Creating Feminist Infrastructures in the Digital Humanities. Krakow http://dh2016.adho.org/abstracts/233 (accessed 10 March 2017).

Gooding, P. (2016). Exploring the information behaviour of users of Welsh Newspapers Online through web log analysis. *Journal of Documentation*, **72**(2): 232–46.

Kua, E. (2008). Non-Western Languages and Literatures in the Dewey Decimal Classification Scheme. *Libri*, **54**(4): 256–265 doi:10.1515/LIBR.2004.256.

Mai, J. (2010). Classification in a social world: bias and trust. *Journal of Documentation*, **66**(5): 627–42 doi:10.1108/00220411011066763.

Nicholas, D., Jamali, H. R. and Huntington, P. (2005). The use and users of scholarly e-journals: a review of log analysis studies. *Aslib Proceedings*, **57**(6): 554–71.

Palmer, C. L. and Cragin, M. H. (2008). Scholarship and disciplinary practices. *Annual Review of Information Science and Technology*, **42**(1): 163–212 doi:10.1002/aris.2008.1440420112.

Ross, C. and Terras, M. (2011). Scholarly Information-Seeking Behaviour in the British Museum Online Collection. Philadelphia, PA http://www.museumsandtheweb.com/mw2011/papers/scholarly_information_seeking_behaviour_in_the (accessed 10 October 2012).

Sinn, D. and Soares, N. (2014). Historian's Use of Digital Archival Collections: The Web, Historical Scholarship, and Archival Research. *Journal of the Association for Information Science and Technology*, **Online First** http://onlinelibrary.wiley.com/doi/10.1002/asi.23091/abstract (accessed 26 June 2014).

Stone, S. (1982). Humanities Scholars: Information Needs and Uses. *Journal of Documentation*, **38**(4): 292–313.

sub-subroutine (2015). Visualising the distribution of human knowledge in the Dewey Decimal System *Sub-Subroutine* http://www.subsubroutine.com/sub-subroutine/2014/12/29/dewey-decimal-system-and-books-filed-therein (accessed 11 November 2018).

Talja, S. and Maula, H. (2003). Reasons for the use and non-use of electronic journals and databases: A domain analytic study in four scholarly disciplines. *Journal of Documentation*, **59**(6): 673–91 doi:10.1108/00220410310506312.

Unsworth, J. (2000). Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might our Tools Reflect This?. King's College London http://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html (accessed 5 August 2014).

Warwick, C., Terras, M., Huntington, P. and Pappa, N. (2008). If You Build It Will They Come? The LAIRAH Study: Quantifying the Use of Online Resources in the Arts and Humanities. *Literary and Linguistic Computing*, **23**(1): 85–102.

Weingart, S. (2014). The moral role of DH in a data-driven world. Lawrence, Kansas http://www.scottbot.net/HIAL/?p=40944 (accessed 2 November 2018).

(2013). *The Legal Deposit Libraries (Non-Print Works) Regulations 2013*. http://www.legislation.gov.uk/uksi/2013/777/contents/made (accessed 15 August 2013).