Scaling Up Bibliographic Data Science

Mikko Tolonen[1][0000−0003−2892−8911], Jani Marjanen[1][0000−0002−3085−4862], Hege Roivainen[1][0000−0002−0489−3278], and Leo Lahti[2][0000−0001−5537−637X]

[1] University of Helsinki, Helsinki, Finland
{first.last}@uni-heidelberg.de
[2] University of Turku, Finland
leo.lahti@iki.fi
http://www.iki.fi/Leo.Lahti

Abstract. Bibliographic data science is an emerging research paradigm in digital humanities. It aims at systematic quantification of the trends in knowledge production based on large-scale analysis of bibliographic metadata collections and the methods of modern data science. Compared to the earlier related attempts in book history and sociology of literature, advances in data processing and quality control are now making it possible for the first time to scale up the analysis to millions of print products while at the same time paying attention to data quality, representativity and completeness. This provides a new quantitative method that can support the analysis of classical research questions in intellectual history. Here, we discuss the methodological challenges that we have encountered in such studies and how to scale up the solutions based on collaborative research efforts.

Keywords: Library catalogues · Data science ecosystem · Publishing history · Open science

1 Introduction

Bibliographic data collections contain a vast array of data on publishing activities, and they have been traditionally used as a tool for information retrieval. A systematic analysis of bibliographic catalogues can generate rich information on historical patterns in knowledge production, and their research potential has been debated for at least half centuries now [14]. Large-scale integration of data from bibliographies and supporting information sources across long time periods, geographical areas, and genres has the potential to shed new light on classical hypotheses as well as to uncover previously overlooked historical trends. Such analyses depend critically not only on data quality and completeness but also on understanding the historical context. Hence, seamless collaboration between data scientists and historians is crucial for obtaining robust conclusions and for developing efficient research methods for such analysis. Whereas earlier studies in analytical bibliography and related fields have discussed these opportunities, emphasized the role of quantification [12, 13, 17, 6, 11, 4] and charted the long-term developments in the history in books [5, 1, 2, 7], systematic quantitative research

use of large bibliographic metadata collections has proven to be challenging. We recently proposed the concept of bibliographic data science (BDS) [9] and provided the first case studies to demonstrate the research potential of this approach. In this paper we complement the previous work by discussing practical solutions to scaling up these efforts in order to integrate data across dozens of bibliographic collections that comprise altogether to millions of print products over several centuries, genres, and languages.

2 Methodological aspects

Bibliographic data science (BDS) [9] facilitates research use of library catalogues by developing systematic quantitative methods to ensure data quality, and support analysis and interpretation. It has been suggested that scaling up bibliographic data science to cover dozens of catalogues and millions of entries could substantially expand the depth and scope of such studies but this depends on our ability to develop scalable solutions for reliable data harmonization and analysis. Such efforts can remarkably benefit from the latest developments in open data science (see e.g. [8, 3]). Compared to other data science projects with related challenges, unique to our efforts is the central role of historical interpretation and the necessity of incorporating prior knowledge on the data collection processes which may introduce unexpected biases in the analyses. Here, we briefly discuss some of the key elements that can help to scale up bibliographic data science. These include automation, standardization, use of supporting data sources, quality monitoring, machine learning, and open collaboration models.

Harmonization of the original records is the first step in facilitating reliable research use. Bibliographic metadata is often manually entered in the databases, and seldom sufficiently standardized and readily amenable to quantitative analysis. Biases, inaccuracies, gaps, and varying standards and languages pose challenges for data integration both within and across catalogues. The heterogeneity of the data fields, spanning from time intervals to persons, physical dimensions, or geographical locations is also a notable challenge. The scale of these issues greatly exceeds our ability to manually verify and correct the entries. Essential for this undertaking is to move towards automated analysis workflows that enable standardized treatment of the data and efficient implementation of similar workflows across multiple catalogues and iterative data corrections.

The integration of data across catalogues enables the analysis of publishing activity beyond what is accessible by the use of individual national bibliographies alone, as we have recently suggested in [16]. Since all bibliographies follow the conventions that are described in the MARC cataloguing format, we have been able to apply largely identical processing across all collections, thus facilitating transparent and scalable data processing. We have extensively harmonized selected fields of the Finnish and Swedish National Bibliographies (FNB and SNB, respectively), the English Short-Title Catalogue (ESTC), and the Heritage of the Printed Book database (HPBD). Altogether, these four bibliographies cover over 6 million entries of print products printed in Europe and elsewhere, and 2.64 mil-

lion harmonized entries in the period 1500-1800, ranging from the 16365 entries in the FNB to 2.1 million entries in HPBD, which is a compilation of 45 smaller, mostly national, bibliographies[3].

Standardization and reproducibility can benefit from the design of dedicated software packages and unit tests. Hence, we have implemented systematic algorithmic approaches that help to assess and improve data reliability in a semi-automated fashion. These workflows combine various procedures to remove spelling errors, disambiguate and standardize terms, augment missing values, and incorporate manually curated information such as known pseudonymes or synonymous entries. Some steps are straightforward, including removal of extra periods or spaces; in other cases, custom algorithms have to be developed. One example is the page count information, which follows a specific MARC notation[4] that separates paginations in different parts of the document, such as preface, contents, figures and tables, and in separate volumes. Specifically designed algorithms are necessary for scalable interpretation of such information.

Remarkable portions of information is typically missing in library catalogues but can be readily augmented based on information that is already contained elsewhere in the catalogue, or based on external information sources. For instance, missing country information can be filled in when the publication place is available and uniquely mappable to a country; and exact physical dimensions of a document can be often inferred with a good accuracy based on the gatherings information that is more frequently available. In such cases, information is however retained whether the value is based on the original entries, or later estimated based on other information. We have also complemented the original records with entirely new derived fields, such as print area, which quantifies the paper consumption per unique physical copy of a document, or title. The combined print area across all unique titles then reflects the overall breadth of printing activities per time period, genre, or geographic location, rather than the mere volume of print products. Hence, it complements total paper consumption which quantifies the overall volume of print products but does not consider the uniqueness or diversity across titles. Estimating the overall paper consumption is more difficult as it requires information on print run sizes and this is often not readily available. The estimated print run size of one thousand copies is often used for early-modern books but notable variations over time and geography have been reported. The print area and paper consumption can hence provide complementary ways to assess trends in knowledge production.

Monitoring the quality of the harmonization process is an important part of such efforts. We are routinely generating conversion tables that show how the original raw entries have been converted into the final harmonized versions, list of common entries in each field (e.g. authors, publishers, languages), and statistical summaries such as average document dimensions by format, histograms of publication years and author life spans, or changes in gender distribution

[3] https://www.cerl.org/resources/hpb/content

[4] see the Library of Congress web document https://www.loc.gov/marc/bibliographic/ for the full description of the MARC21 format.

over time. Such overviews have been invaluable in spotting unexpected events that may in turn facilitate the detection of remaining inaccuracies or biases in the harmonized data sets. The quality monitoring has been greatly facilitated by such reproducible overviews, and the most up-to-date summaries for each bibliographic catalogue that we are working on can be accessed via Helsinki Computational History Group website[5]. By openly sharing the summaries and algorithms, we are aiming to improve the transparency, reliability, and overall quality of our work by providing the broader audience with the means to detect and report potential inaccuracies.

Advances in machine learning and artificial intelligence are providing further means to scale up the analysis. We are routinely utilizing methods from natural language processing, feature selection, clustering, and classification to facilitate duplicate identification and quality monitoring. Information in the publisher field, for instance, often contains lengthy verbal explanations that have to be subjected to named entity recognition. Moreover, detecting alternative spellings for author or publisher names can be greatly accelerated by string distance matching algorithms that can rank potential duplicates and facilitate semi-automated data curation. Page number estimation is a prominent example of the potential of this approach, as the original text entries can be converted into a single well-defined number, which is the final page count estimate. The overall accuracy of such fully automated estimates can be quantified to a great extent by examining the most common conversions and representative sets of random examples. Future developments could take increasing advantage of such statistical techniques in order to reduce the need for human input, thus improving the overall scalability of data harmonization. The already curated data sets provide ample training material for supervised machine learning techniques.

Finally, the research community can benefit from open sharing of the data and algorithms. The lack of open data availability is forming a major bottleneck for collaborative development of bibliographic data science but this might be gradually changing. The National Library of Finland, for instance, recently released the complete MARC entries of the FNB[6] under an open data license that allows the modification, reuse, and sharing of derivative versions. We have made the key algorithms for harmonization and analysis openly available in the bibliographica R package[7], and our harmonized versions of the FNB data set can be accessed via Helsinki Computational History Group website. The harmonized data sets can be further verified, investigated, and enriched by others, and they could be integrated into Linked Open Data and other popular formats in order to utilize the vast pool of existing software tools. Combining such large-scale harmonization with existing data management infrastructures could open up new doors for research on national bibliographies.

[5] https://www.helsinki.fi/en/researchgroups/computational-history
[6] http://data.nationallibrary.fi/
[7] https://github.com/COMHIS/bibliographica

3 Emerging applications

Our recent analyses based on four large bibliographies provides examples of the research potential of this approach [9, 16]. One example is the analysis of the long-term development of book formats and languages across Europe, which reflects changes in public communication. We have reported that on a general European level the rise of the octavo format is particularly strong during the eighteenth century, and supported by the data in all four catalogues, where octavo holds the largest share of the print area by the end of the eighteenth century. All four metadata collections that we have analysed show a steadily declining, parallel trend in the share of publications in Latin in the period 1500-1800. Such observations can highlight material aspects of vernacularization in the early modern period and reflect European-wide transformations that took place predominantly during the hand-press era. Moreover, the analyses can also highlight local variations in the publication profiles of individual European cities [9]. Therefore joint analysis and comparison of multiple catalogues can be useful also in terms of assessing the historical representativity of the data, thus demonstrating the value of bibliographic data science and paving the way for new research and guidelines for future data integration in this field.

4 Conclusion

We have conceptualized a new approach, bibliographic data science, to expand the research potential of bibliographic records. This derives from the already established field of data science and associates this general paradigm specifically with quantitative analysis of bibliographic metadata and related information sources. While having a specific scope, BDS is opening up pragmatically oriented and substantial new research opportunities in the digital humanities.

Drawing valid conclusions critically depends on data quality, representativity and completeness. Automation and quality control are essential when the data collections may contain information on millions of documents, and the overall data science ecosystem integrates a number of distinct workflows that are dedicated to harmonizing specific subsets of the data. We have indicated how specifically tailored open data analytical ecosystems can help to address this challenge. Our approach has potential for wider implementation in related studies, and provides guidelines for more extensive integration of national collections.

Our future work envisions continued harmonization and data integration for the HPBD as well as further, related data resources such as the Universal Short Title Catalogue.[8], in order to expand the study to cover public communication more broadly. We have incorporated best practices and tools from data science, such as unit tests, tidy data [18] and reproducible workflows [19]. Future developments could take increasing advantage of machine learning in order to reduce the need for human input, thus improving the overall scalability of data harmonization. When combined with a proper quality control, such approaches can have

[8] https://ustc.ac.uk/

potential for wider implementation in related studies in the digital humanities. As we have extracted and harmonized publisher information from imprints from ESTC and FNB [15], it is possible to connect that data to full-text collections such as the ECCO, and to study how the materiality of printing is related to developments in newspapers [10]. Modern statistical techniques are essential not only the harmonization and quality control of the data but also in investigating and characterizing the overall spatio-temporal trends, networks, and dynamics in knowledge production. Taking full advantage of the developments in open sharing of research data and analysis methods can support collaborative and cumulative research efforts. Automated harmonization can enhance the overall reliability and commensurability between independently maintained metadata collections, thus complementing linked open data and other technologies that primarily focus on data management and distribution. Hence, bibliographic data science can help to fill an important gap in the field by aiming to significantly improve the quality and reliability of the currently available bibliographic records.

References

1. Baten, J., van Zanden, J.L.: Book production and the onset of modern economic growth. Journal of Economic Growth **13**(3), 217–235 (2008). https://doi.org/10.1007/s10887-008-9031-9
2. Bell, M., Barnard, J.: Provisional Count of STC Titles, 14751640. Publishing History **31**(1), 4764 (1992)
3. Borgman, C.L.: Big data, little data, no data : scholarship in the networked world. The MIT Press, Cambridge, Massachusetts; London, England (2015)
4. Bozzolo, C., Ornato, E.: Pour une histoire du livre manuscrit au Moyen Age : trois essais de codicologie quantitative. Equipe de recherche sur l'humanisme francais des XIVe et XVe siecles, Editions du Centre national de la recherche scientifique, Paris (1980)
5. Buringh, E., Zanden, J.L.V.: Charting the "Rise of the West": Manuscripts and Printed Books in Europe A Long-Term Perspective from the Sixth through Eighteenth Centuries. The Journal of Economic History **69**(02), 409 (2009). https://doi.org/10.1017/s0022050709000837
6. Giesecke, M.: Der Buchdruck in der frhen Neuzeit : eine historische Fallstudie uber die Durchsetzung neuer Informations- und Kommunikationstechnologien. Suhrkamp, Frankfurt am Main (1991)
7. Horstbøll, H.: Menigmands medie: det folkelige bogtryk i Danmark 1500-1840: en kulturhistorisk undersøgelse. Danish humanist texts and studies, volume 19, Det Kongelige Bibliotek & Museum Tusculanum, Copenhagen (1999)
8. Lahti, L.: Open data science. In: Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191. vol. 11191. Springer, India (October 2018), conference proceedings.
9. Lahti, L., Marjanen, J., Roivainen, H., Tolonen, M.: Bibliographic data science and the history of the book (c. 15001800). Cataloging & Classification Quarterly pp. 1–19 (January 2019). https://doi.org/10.1080/01639374.2018.1543747, special issue.

10. Marjanen, J., Vaara, V., Kanner, A., Roivainen, H., Mäkelä, E., Lahti, L., Tolonen, M.: Analysing the language, location and form of newspapers in finland, 1771-1910. Tech. rep., Digital Humanities in the Nordics, Gothenburg (2017), conference abstract.
11. Neddermeyer, U.: Von der Handschrift zum gedruckten Buch : Schriftlichkeit und Leseinteresse im Mittelalter und in der fruhen Neuzeit : quantitative und qualitative Aspekte. Buchwissenschaftliche Beitrge aus dem Deutschen Bucharchiv Mnchen, Harrassowitz, Wiesbaden (1998), 1: Text ; 2: Anlagen.
12. Suarez, M.F.: Towards a bibliometric analysis of the surviving record 1701–1800. In: Suarez, M.F., Turner, M.L. (eds.) The Cambridge History of the Book in Britain, pp. 37–65. Cambridge University Press (2009). https://doi.org/10.1017/chol9780521810173.003
13. Suarez SJ, M.F.: Book history from descriptive bibliographies. In: Howsam, L. (ed.) The Cambridge Companion to the History of the Book, pp. 199–218. Cambridge University Press (2014). https://doi.org/10.1017/cco9781139152242.015
14. Tanselle, G.T.: Bibliography and science. Studies in Bibliography **27**, 55–90 (1974)
15. Tolonen, M., Lahti, L., Roivainen, H., Ilomki, N.: Printing in a Periphery: a Quantitative Study of Finnish Knowledge Production, 1640-1828. In: Digital Humanities 2016: Conference Abstracts. pp. 383–385. Jagiellonian University & Pedagogical University, Krakw (2016)
16. Tolonen, M., Lahti, L., Roivainen, H., Marjanen, J.: A quantitative approach to book-printing in sweden and finland, 16401828. Historical Methods: A Journal of Quantitative and Interdisciplinary History pp. 1–22 (2018). https://doi.org/10.1080/01615440.2018.1526657
17. Weedon, A.: The Uses of Quantification. In: Eliot, S., Rose, J. (eds.) A Companion to the History of the Book, pp. 33–49. Blackwell Publishing Ltd, London (2008). https://doi.org/10.1002/9780470690949.ch3
18. Wickham, H.: Tidy data. Journal of Statistical Software **59**(10), 1–23 (2014). https://doi.org/10.18637/jss.v059.i10
19. Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., Teal, T.K.: Good enough practices in scientific computing. PLoS Computational Biology **13**(6), e1005510 (2017). https://doi.org/10.1371/journal.pcbi.1005510