Best Practices in Bibliographic Data Science

Leo Lahti[1], Ville Vaara[2], Jani Marjanen[2], and Mikko Tolonen[2]

[1] University of Turku, Finland
[2] University of Helsinki, Helsinki, Finland

Abstract. Bibliographic data science aims to quantify historical trends in knowledge production based on the rich information content in library catalogue metadata collections. Compared to the earlier attempts in book history, advances in data science are now making it possible to automate remarkable portions of such analyses, while maintaining or improving data quality and completeness. Such quantitative approaches can support the analysis of classical questions in intellectual history. Here, we discuss best practices in this emerging research field.

1 Introduction

Large-scale integration and analysis of bibliographic data collections and supporting information sources across time, geography, or genre could shed new light on classical hypotheses and uncover overlooked historical trends. This quantitative research potential has been long recognized [1–5]. Systematic large-scale research use of bibliographic collections has proven to be challenging, however.

Bibliographic data science (BDS) [6] develops systematic methods for the harmonization, analysis and interpretation of bibliographic metadata collections. Scaling up bibliographic data science can remarkably benefit from the developments in open data science [7–9]. Automated harmonization can enhance the quality and commensurability between metadata collections, thus complementing linked open data and other technologies that focus on data management and distribution. Unique to our efforts is the focus on historical interpretation and the necessity of incorporating prior knowledge on the data collection processes which may introduce biases in the analyses. However, ensuring data quality and completeness are critical for research, and opening up the research workflows can support collaboration across institutional and national borders.

In order to address these challenges, we have recently proposed the concept of bibliographic data science (BDS) [6] and provided the first case studies to demonstrate its research potential. Here, we discuss challenges and best practices that can facilitate cumulative research efforts in this field.

2 Best practices

2.1 Reproducible data harmonization and analysis

Bibliographic metadata is often manually entered in the databases, and seldom sufficiently standardized for systematic quantitative analysis. Harmonization of the raw data entries is the first step towards reliable research use. Biases,

gaps, and varying standards pose challenges for data integration both within and across catalogues. Heterogeneity of the data fields, from time intervals to persons, physical dimensions, or geographical locations form challenges for analysis [7]. The metadata fields can often have complex dependencies, and harmonization of one field may influence the harmonization and enrichment of the other.

In order to efficiently manage these multi-layered data harmonization processes, we have implemented a data science ecosystem that integrates various individual workflows (Figure 1), incorporating pragmatic approaches from data science [10, 11]. These workflows remove spelling errors, disambiguate and standardize terms, augment missing values, and incorporate manually curated information on pseudonymes, duplicate entries, and other information types. Reproducible workflows support transparent data processing which can be efficiently monitored and improved over time. This can further benefit from the design of dedicated software packages[3]. We have used these tools to extensively harmonize selected fields of the Finnish and Swedish National Bibliographies (FNB and SNB, respectively), the English Short-Title Catalogue (ESTC), and Heritage of the Printed Book database (HPBD). Altogether, these collections cover over 6 million entries of print products printed in Europe and elsewhere. Since all bibliographies are MARC compatible[4], we can apply largely identical processing across all collections, thus facilitating transparent and scalable data processing [6, 12, 7].

2.2 Semi-automated data curation

The scale of data harmonization greatly exceeds our ability to manually correct and verify individual entries. Automated workflows can be used to standardize data treatment across multiple catalogues, allowing iterative corrections. In addition to inaccurate or erroneous entries, duplicates and missing information can induce systematic bias in the analyses, in particular when they are unevenly distributed. Such biases can be detected by systematic quality monitoring (Figure 2). Missing information can be often augmented based on other information sources. For instance, missing country information can be inferred when the publication place is known, and information on authors can be found from historical databases.

Whereas manual verification remains a key component in data curation, this can be greatly facilitated by automation. We have implemented systematic approaches to assess and improve data reliability in a semi-automated fashion based on unit tests, outlier detection, and external verification. Automatically generated summary reports of the harmonized data are a key part of this process, and can be accessed for the different bibliograhic catalogues, such as the ESTC and FNB, through the homepage of Helsinki Computational History group. An example of this is the comparison of page counts between a subset of the ESTC

[3] https://github.com/COMHIS/bibliographica
[4] Library of Congress web document https://www.loc.gov/marc/bibliographic/

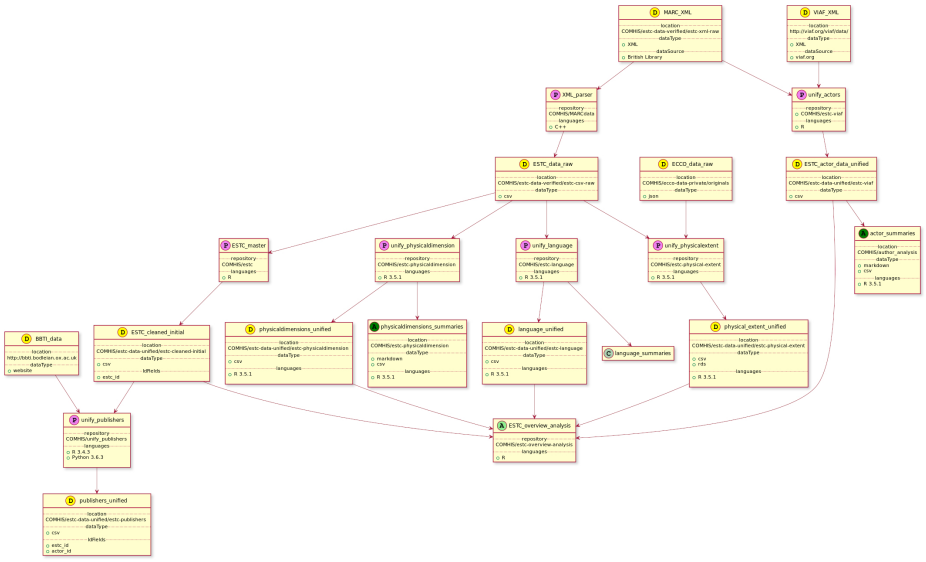

Fig. 1. Dependency chart describing the multi-layered data harmonization process of the English Short Title Catalogue.

entries from our workflow, and a distinct database, the Eighteenth Century Collections Online (Figure 3). Furthermore, we generate conversion tables that show how common entries have been harmonized. By visualizing complementary aspects of the data such as document dimensions, publication years, author life spans, or gender distributions based on timelines, scatterplots, histograms and heatmaps we can obtain further insights to potential inconsistencies (Figure 4). Such overviews can be invaluable for the detection of inaccuracies or biases in the harmonized data sets.

Moreover, data integration across catalogues enables the detection of robust patterns that are supported by multiple information sources. Joint analysis can be useful in terms of assessing the historical representativity, thus paving the way for future data integration and analysis. Our recent analyses provide examples of the research potential of this approach [6, 12]. For instance, all metadata collections that we have analysed show a systematic rise of the octavo format during the eighteenth century and a parallel, steadily declining trend in Latin publications towards the eighteenth century [6].

2.3 Open science and collaboration

Academic research can benefit enormously from open sharing of data and algorithms [13]. There are differences between research fields, however. Open data sharing in the humanities is still less well developed than in natural sciences but this might be gradually changing. In Finland, for instance, The National Library

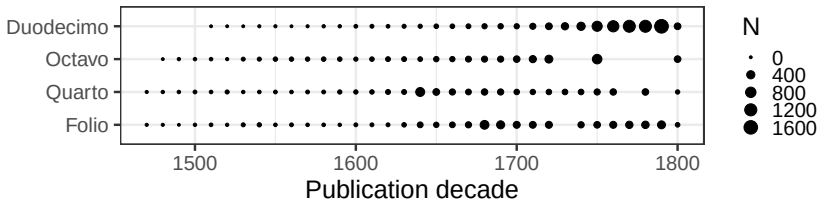

Fig. 2. Visualization of missing page count information by decade (horizontal axis) for different book formats (vertical axis) reveals systematic biases in data availability in the ESTC.

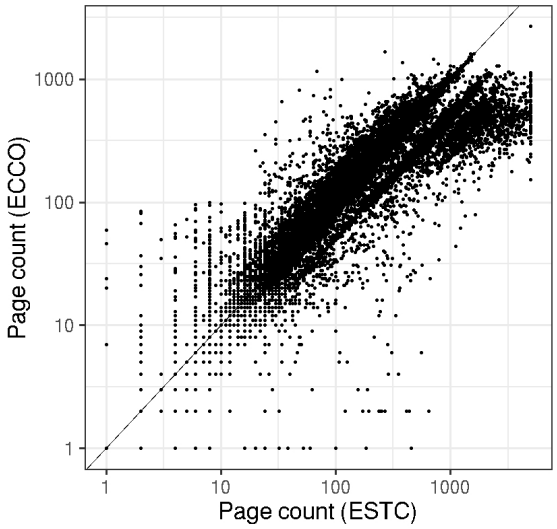

Fig. 3. Comparison of page count estimates to an external validation set. The English Short Title Catalogue (ESTC) includes 175727 documents that have a page count estimate and can be matched with the Eighteenth Century Collections Online (ECCO) database. The ECCO database contains page count information for the same works. The Spearman correlation between the two sources is $\rho = 0.99$. The comparison quantifies the quality of data harmonization and can potentially highlight issues that would benefit from further improvements in the data harmonization workflow.

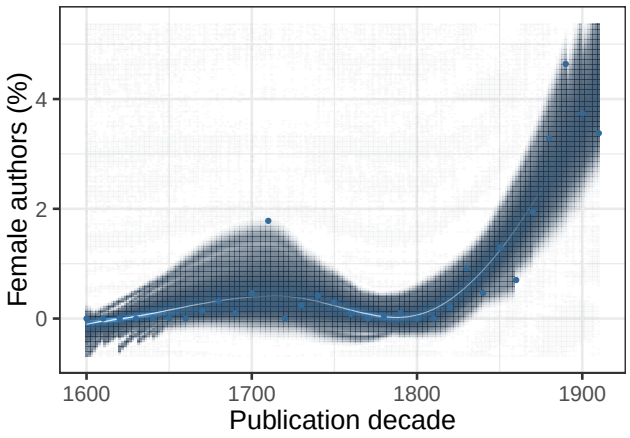

Fig. 4. Proportion of female authors in the Finnish national bibliography (FNB) in the period 1600-1900. The visualization can also highlight unexpected patterns in the data, such as outlier points and the temporary increase in the early 18th century, which may indicate inaccuracies in the author harmonization and gender identification algorithms.

has recently released the FNB under an explicit open data license[5]. This has allowed us to generate and share further, harmonized versions that can be accessed via Helsinki Computational History Group website[6]. Combining such large-scale harmonization with existing infrastructures could open up new opportunities for research.

Open sharing of research data and algorithms can facilitate collaboration across institutional and national borders and increases transparency and efficiency in research. By making our key algorithms, software packages, and workflows available under open licenses [14] we are allowing others to search, detect and report inaccuracies, or make further corrections and utilize these resources in independent research and methods development.

2.4 Contributions and interaction between fields

The research methods discussed in this paper contributes to traditionally distinct fields. The study is motivated by classical questions in book history, intellectual history, and early modern history. In addition to describing broad trends in knowledge production, it is becoming possible to provide exact large-scale quantification of historical trends using library metadata catalogues. This can be contrasted with the existing knowledge in order to confirm earlier hypotheses and as well as to highlight trends that so far received less attention. The new quantitative techniques bring up new historical information and evidence,

[5] http://data.nationallibrary.fi/
[6] https://www.helsinki.fi/en/researchgroups/computational-history

thus complementing more traditional qualitative methods in history and related fields. At the same time, the work contributes to methodological research from linguistics and natural language processing to machine learning, data science, and interaction design. Quantitative methods can support data harmonization and analysis and integration of the metadata collections with complementary data types such as full texts and geographical information. The new applications can inspire further research in the methodological fields.

3 Conclusion

This short paper has reviewed current and emerging best practices in bibliographic data science. This is an emerging research paradigm in the digital humanities, which focuses on the research use of bibliographic metadata and potentially other related data types such as full text collections and complementary information on persons, organizations and places. Whereas this overall research theme is opening up opportunities to study a variety of traditional research questions in history, sociolinguistics or related fields from a new angle, the focus on the present work has been in demonstrating the challenges for methodological research [6, 12, 15]. There are a number of technical research questions on how to optimally design, choose, implement, validate, or utilize the research algorithms in general and in the specific context our application domain (see e.g. [8]); a full discussion of these aspects is beyond the scope of this short paper, however. Specifically relevant for our application are the questions on how to combine qualitative and quantitative elements in a reliable and unbiased manner; or at least how to identify and deal with the possible biases, inaccuracies, and gaps in the research data and analysis process. Whereas the concept of bibliographic data science has been proposed only recently [6], it builds on the existing traditions of bibliographic research [1–5] and open data science (see e.g. [14, 16, **?**]). Bibliographic data science is expanding the research potential of large-scale library catalogues. Adaptations of this method can be used in other related fields, such as the study of materiality in newspapers [17]. Automation and quality control are essential when the data collections cover millions of documents. Advances in machine learning and artificial intelligence are now providing further means to scale up the analysis as the already curated data sets provide ample training material for supervised machine learning techniques. Open sharing of research data and analysis methods can support collaborative and cumulative research efforts. We have demonstrated how specifically tailored open data analytical ecosystems can help to address these challenges.

Acknowledgments We would like to thank members of Helsinki Computational History group for their contributions to data harmonization, Hege Roivainen for his support in the analysis, and Eetu Mäkelä for providing the ECCO page count information. This work has been supported by Academy of Finland (decision 293316).

References

1. Tanselle GT (1974) Bibliography and science. Studies in Bibliography 27: 55–90.
2. Giesecke M (1991) Der Buchdruck in der frühen Neuzeit: eine historische Fallstudie uber die Durchsetzung neuer Informations- und Kommunikationstechnologien. Suhrkamp, Frankfurt am Main.
3. Bozzolo C & Ornato E (1980) 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.
4. Bell M & Barnard J (1992) Provisional Count of STC Titles, 1475-1640. Publishing History 31(1): 4764.
5. Horstbøll H (1999) 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.
6. Lahti L, Marjanen J, Roivainen H & Tolonen M (2019) Bibliographic data science and the history of the book (c. 1500-1800). Cataloging & Classification Quarterly pp. 1–19. Special issue.
7. Tolonen M, Marjanen J, Roivainen H & Lahti L (2019) Scaling up bibliographic data science. In: Proceedings of the Digital Humanities in the Nordics (DHN2019).
8. Lahti L (2018) Open data science. In: Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191.
9. Borgman CL (2015) Big data, little data, no data: scholarship in the networked world. The MIT Press, Cambridge, Massachusetts; London, England.
10. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L & Teal TK (2017) Good enough practices in scientific computing. PLoS Computational Biology 13(6): e1005510.
11. Wickham H (2014) Tidy data. Journal of Statistical Software 59(10): 1–23.
12. Tolonen M, Lahti L, Roivainen H & Marjanen J (2018) A quantitative approach to book-printing in Sweden and Finland, 1640-1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History pp. 1–22.
13. Morin A, Urban J, Adams P, Foster I, Sali A, Baker D & Sliz P (2012) Research priorities. shining light into black boxes. Science 336(6078): 159–60.
14. Morin A, Urban J & Sliz P (2012) A Quick Guide to Software Licensing for the Scientist-Programmer. PLoS Computational Biology 8(7): e1002598.
15. Lahti L, Ilomäki N & Tolonen M (2015) A Quantitative Study of History in the English Short-Title Catalogue (ESTC) 1470-1800. LIBER Quarterly 25(2): 87–116.
16. Ioannidis J (2014) How to make more published research true. PLoS Medicine 11(10): e1001747.
17. Marjanen J, Vaara V, Kanner A, Roivainen H, Mäkelä E, Lahti L & Tolonen M (2017) Analysing the language, location and form of newspapers in finland, 1771-1910. Technical report, Digital Humanities in the Nordics, Gothenburg. Conference abstract.