The Emerging Paradigm of Bibliographic Data Science

Introduction

Library catalogues contain rich, albeit potentially incomplete information on historical trends and shifts in knowledge production. Their research potential has been debated for more than 50 years (Tanselle, 1974). Large-scale harmonization and analysis of these data collections has provided new ways to investigate classical research hypotheses in book history and intellectual history. Whereas the potential biases in the data, arising for instance from variations in data collection practices over time and place, have to be carefully considered in the analysis, a data-driven approach provides a uniquely large-scale view and a supportive role in research. Despite related earlier work (Buringh and van Zanden, 2009; Bell and Barnard, 1992), a systematic research use of bibliographic metadata has proven to be challenging, however. The lack of scalable solutions for improving and verifying data quality and completeness has been a major bottleneck for large-scale analysis. Our team has recently proposed the concept of bibliographic data science in order to overcome some of these challenges. Here we provide an overview of the ongoing attempts to develop a broader research line that focuses on the development of targeted analysis methods in this research area, and specifically demonstrate the advantages of fully open bibliographic data science. A number of specific case studies and applications of this approach are included in the proceedings of this conference.

Data and methods

Bibliographic data science (Lahti et al., 2019) is an emerging paradigm in the digital humanities. It aims to improve the overall data reliability and completeness through systematic harmonization, error correction, and enrichment of missing information, thus greatly enhancing the research potential of bibliographic collections. It derives from the paradigms of open science and data science (see e.g. Lahti et al., 2015; Tolonen et al., 2016; Lahti 2018b), and incorporates best practices from these fields, including reproducible analysis, open source code, and open collaboration models. We have recently integrated metadata across four large bibliographies and altogether 2.64 million harmonized entries in the period c. 1500–1800 (Tolonen et al., 2019) from the Finnish and Swedish National Bibliographies, the English Short-Title Catalogue, and the Heritage of the Printed Book database. Compared to the earlier efforts, our automated approach is uniquely scalable and comprehensive in terms of data integration and quality monitoring. Furthermore, it is combined with systematic expert curation and research use that allow us to detect shortcomings and inconsistencies that are historically relevant but challenging to observe by automated means. As such, our newly implemented methods exemplify the application of augmented intelligence in digital humanities data curation and analysis.

Case studies

Large-scale bibliographies are seldom available as open data. This has formed a bottleneck for the development of new data processing and analysis methods as their usability and value is very limited and restricted to only those research groups who have access to the same or similar data collections. This is in contrast to some other data-intensive fields such as bioinformatics, where considerable data resources are openly shared by national and international research organizations, and algorithmic tools to access and utilize them are being routinely developed and shared widely within the research community. The National

Library of Finland has, however, released The Finnish National Bibliography in an openly licensed, machine-readable format (National Library of Finland, 2017). This combination of open data and open analysis workflows allows us to demonstrate the opportunities of fully open bibliographic data science. We have previously estimated the long-term development of book formats, which reflects shifts in reading habits and public communication over time (Lahti et al. 2015; Lahti et al. 2019; Tolonen et al. 2019). One example is the observed changes is the rise of the octavo format, which supersedes other printing formats during the eighteenth century, in parallel with a systematic decline in the use of Latin and a growing share of published books printed in vernacular languages. Here we complement such case studies by demonstrating the challenges and opportunities in opening the complete analysis workflows, and show how this establishes the overall methodological basis for the more specific case studies that are being presented in this conference.

Conclusion

Bibliographic data science is renewing research in digital humanities in general, and in book history in particular. Related data harmonization efforts include for instance the Collections as Data project (Padilla et al. 2019), which has promoted generic research use of data collections in digital humanities and related fields. In contrast, our work explicitly focuses on the specific research area of early-modern knowledge production and intellectual history. Our focus is therefore more specific than in generic data science projects. The scalability of the work over additional metadata fields and different time periods can pose remarkable further challenges which could be partially addressed by focusing on specific topics of interest, such as variations in authors, language use, publishing networks or document materiality, as is often the case in pragmatic research projects. Our vision includes combining the harmonized metadata with full-text collections such as the ECCO, and studying how the materiality of printing is related to developments in newspapers (Marjanen et al., 2017). When combined with a proper quality control, such data-driven approaches have potential for wider implementation in related studies in the digital humanities. Hence, the contribution of this work is not merely in the development or application of new algorithms or exploration techniques, but in demonstrating their wider potential in advancing the methodological basis of the field.

Funding

This work was supported by the Academy of Finland [grant number 293316].

References