Bibliographic Data Science

Leo Lahti, Jani Marjanen, Hege Roivainen, Mikko Tolonen. Helsinki Computational History Group.

The use of library catalogues as research material has been rapidly increasing in the recent years. Library catalogues contain rich metadata on knowledge production trends, with the potential to verify and complement earlier hypotheses as well as to uncover previously overlooked historical trends. However, obtaining valid conclusions depends on the overall understanding of the historical context as well as data quality. We demonstrate how newly developed bibliographic data science ecosystem can help to overcome the prevailing bottlenecks and has the potential to renew our understanding of knowledge production and the public sphere. In particular, we provide examples based on our recent work based on the integration of data from four large bibliographies that we have extensively harmonized. These include the Finnish and Swedish National Bibliographies, the English Short-Title Catalogue, and the Heritage of the Printed Book database, covering altogether 2.64 million harmonized entries from the investigated period. We characterize patterns in reading habits and changes in the publishing landscape across different genres over time and geography during the period c. 1500-1800. An important aspect of this work is to proceed from data browsing interfaces towards a more systematic computational integration and analysis of the available data resources based on the latest advances in modern data science. As such, our work provides an example on how the use of big data and new quantitative methods can enrich more traditional forms of research in literary studies. 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 overall methodological basis of the field.