Quantitative analysis of public discourse in 1470–1910

Leo Lahti[1], Ville Vaara[2], Jani Marjanen[2], Hege Roivainen[2], Ali Ijaz[2], Simon Hengchen[2], Iiro Tiihonen[2], Tanja Säily[2], Antti Kanner[2], Mark Hill[2], Eetu Mäkelä[2], Mikko Tolonen[2]*

[1]Department of Mathematics and Statistics. University of Turku, Finland <leo.lahti@iki.fi>
[2]Faculty of Arts, University of Helsinki, Finland / [*]Presenting author

Changes in knowledge production reflect and drive societal change. A systematic large-scale analysis of the evolution and spread of printing activity over time and geography can thus highlight key moments of transformation in history. Better understanding of the emergence of the printing press, scientific disciplines, conceptual change, or the timing of these events, can provide new quantitative information, and help to uncover overlooked patterns in knowledge production and public discourse.

In the Helsinki Computational History Group [1], Finland, we are studying public discourse in Finland and Europe during the period 1470–1910 by blending historical and computational approaches [2]. We study books, newspapers and other texts as vehicles of thoughts and to analyse how their development over long time spans reflects social movements. Our focus is on publication patterns in Nordic language areas, where we have integrated national bibliographies to expand the analysis beyond national borders in order to reassess the scope, nature and transnational connections of public discourse. Integration of comprehensive bibliographies and full texts within a robust data analytical framework has allowed us to study how the development of Finnish book, newspaper and journal production compares to the broader Nordic and European trends.

National bibliographies form a rich but remarkably undervalued data resource for historical research, providing comprehensive quantitative insights to the large-scale temporal and spatial dynamics of the evolving publishing landscape. Moreover, the analysis of digitized full text materials can benefit from the broader context that large-scale metadata collections can provide in the study of classical questions, such as change in language, concepts, or social processes in particular geographical areas, disciplines, or over specific time periods. Bibliographies contain comprehensive information on publishing trends, time, place, authors, publishers, genres, languages, physical dimensions, and other information, covering up to millions of print products. Scaling up the analysis by orders of magnitude compared the previous studies helps to overcome occasional inaccuracies as the large historical trends are typically overwhelmingly clear. However, obtaining valid conclusions critically depends on efficient and reliable harmonization, augmentation and enrichment of the raw entries, as well as on understanding the potential domain-specific sources of bias in data collection and availability. Systematic technical biases can arise, for instance, from different archiving conventions or missing data at specific historical periods. The use of digital resources in the humanities is steadily increasing but the field is lacking clear standards on important technical aspects such as data quality and availability, reproducibility of the methods, and transparent reporting of the research process.

We demonstrate that such challenges can be overcome by specifically tailored data analytical ecosystems to clean up, harmonize, integrate, and analyse large-scale bibliographic metadata, full texts, and supplementary data sources. Addressing these issues forms an integral part of the research process. Automation helps to solve two key problems in the field, challenges for reproducibility, which can arise from constant database updates, and the lack of transparency in data analysis. Our programmatic approach allows full flexibility in constructing custom workflows, and harnessing the full potential of modern data analysis and visualization arsenal by borrowing best practices from other fields of data-intensive science.

Whereas the emergence of digital methods and data resources is rapidly bringing new opportunities for studying traditional questions in the humanities, the methodological basis of the field is still shaping up. Vast masses of data provide only a starting point for research and need to be complemented by reliable methods for extracting information, to understand biases and caveats in data availability and analysis, and to combine new observations into the overall body of knowledge. We demonstrate how our open source workflows can harmonize data across large-scale bibliographies in a semi-automated and scalable fashion, with minimal human intervention, thus overcoming the pitfalls of methodological nationalism.

Our results demonstrate how social change and public discourse are intertwined, and how cultural, institutional, legal and technological changes are reflected both in publication metadata and the textual content of the publications. Whereas our main focus is on the analysis of Nordic countries, we demonstrate how our proposed approach generalizes more widely to study the formation of the intellectual landscape and local flavours of early modern and modern Europe. These research questions arise from a core field in the humanities, and our research is carried out in seamless collaboration between data analysts, memory organizations, linguists, and historians to take full advantage of the relevant information. For instance, we demonstrate novel opportunities to characterize the impact of the turn in Finland from Swedish to Russian rule in early nineteenth-century public discourse. We argue that previous historical research has lacked appropriate quantitative tools to take an objective ‘bird’s-eye’ view of these complicated and crucial transformations. Similar analyses that combine library catalogue metadata, full texts, and computational workflows at this scale have not been carried out previously. We argue that this yields a novel, data-driven angle on qualitative key research questions, helps to renew the methodological basis of the field, and showcases the vast opportunities of quantitative analysis of digitized materials in the reinterpretation of key questions in study of history. The applicability of our results and open data analytics extends beyond the scope of this project and will contribute to transforming humanities research.

[1] Leo Lahti, Niko Ilomäki, Mikko Tolonen, 2015. **A Quantitative Study of History in the English Short-Title Catalogue (ESTC) 1470-1800**. *LIBER Quarterly* 25(2):87-116.

[2] Helsinki Computational History Group: https://comhis.github.io/