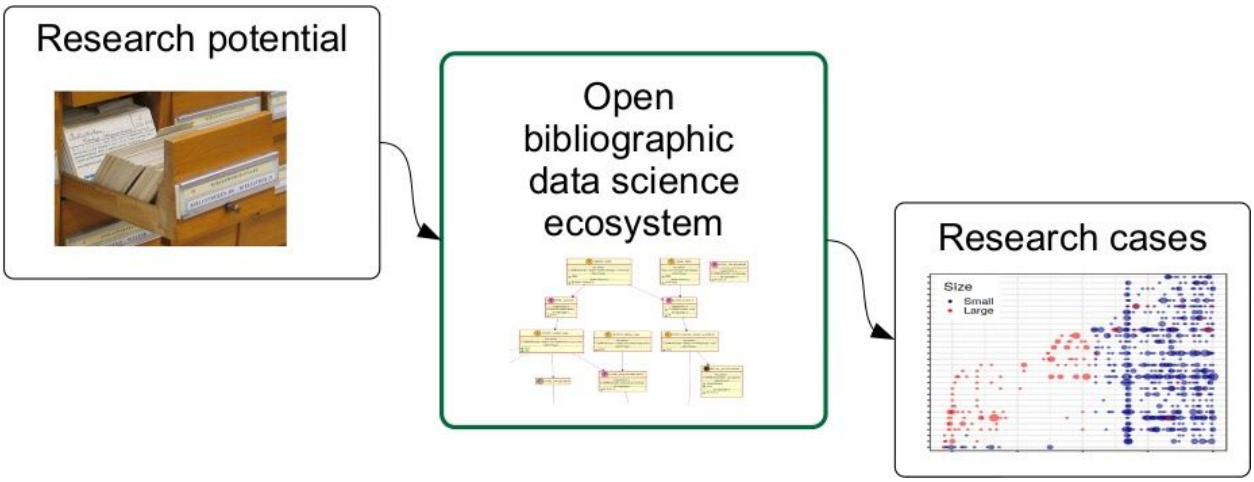

Better Science through Better Data #scidata19

London, Nov 6, 2019. Mikko Tolonen & Leo Lahti

Integrating open science in the humanities: the case of computational history

Data science and computational technologies are now being adopted and are transforming research in classical fields that have traditionally relied heavily on qualitative analysis.

Not long ago it was common for humanities scholars to think that their work does not include research data. The range of sources of a historian, for example, includes original documents in the libraries and archives, secondary literature, and eventually, one's own publications, preferably in the monograph form. When someone started to talk to historians about data and open science, the majority of the community responded with silence. We simply don't have research data, so there is little to talk about.

Humanities research has always aimed for perfection. The most admired products in the field of history, for example, are critical editions where the final word on a matter has been cast. A serious scholar works in the archives and digs up every possible detail about a topic of interest. In this setting, aggregating and using informative albeit incomplete datasets to discover and characterize broad historical trends is frowned upon.

Advances in digitization have brought the concepts of big (and small) data to humanities. Scholars are beginning to realise that perhaps research data has been available after all. However, the idea still remains that a good humanities research project is such that strives for

perfection. If you are using library catalogues to study the overall development of knowledge production, the knowledge of the representativeness and limitations of the data is not enough, many scholars remain unsatisfied if the source of information is not a final product.

With the emphasis on new research methods, the importance of reliable, high-quality research data is easily overlooked. The focus of our own research has been in the analysis of early modern British literature (c. 1500-1800). Library catalogues form a key source of information for this research but the original records are seldom ready for systematic analysis without extensive harmonization, enrichment, and quality control. Instead of thinking that any historical record would be final and critical, we consider these data sets as living entities whose accuracy and interpretation can evolve over time as our collective knowledge accumulates.

We deliberately design our projects not to generate final products. Instead, we construct reproducible ecosystems that accept that there will always be more harmonizing, enriching and linking of sources to be done. Our computational ecosystem that implements these operations is forming a scalable collaboration platform that incorporates our collective and constantly accumulating knowledge base. From the classical humanities perspective this might be difficult to comprehend, but it has induced new strands in humanities research culture that can have long term impact.

More and more research is now done by working in teams with complementary expertise and skills, often initiating from smaller twin projects that extend into larger teams over time. At the same time, the importance of long term planning and the broader research community in developing and openly sharing data and methods has become critical for progress in data-intensive research areas. However, whereas the research community is now actively sharing computational methods and workflows, the lack of open data is forming a severe bottleneck for the development and expansion of quantitatively oriented research in the humanities. Important historical sources, such as library catalogues, may be available only to a handful of dedicated research teams. Open availability of such digital resources could boost quantitatively oriented research in the humanities in the same way as massive genomic databases have been openly released to benefit the global research community.

We have put major efforts in advancing these aspects in the context of intellectual history, in an approach that we call bibliographical data science. Helsinki Computational History Group has brought together expertise from intellectual history, natural language processing, statistical ecology, data science, and related areas. Curated analysis of hundreds of thousands of print products that we have carried out so far has brought up previously uncovered trends in book printing and culture. Similar approaches could be extended to music, museum collections, audiovisual heritage, and other areas where similar systematic records can be collected. Future opportunities that are opened by these efforts will include the use of massive metadata collections to enhance the interpretation of complementary sources, such as full text collections, that would help to overcome earlier shortcomings such as nationally restricted perspectives, with the cross-European perspective that is the big promise of digital humanities.

Given the recent progress and expansion in various fields of computational humanities, we anticipate that the concepts of open science and reproducible research, and collaborative infrastructures that implement such work at a practical level, will have an increasingly central role in this emerging area in the coming years, borrowing ideas and best practices from other, established fields of applied data science.

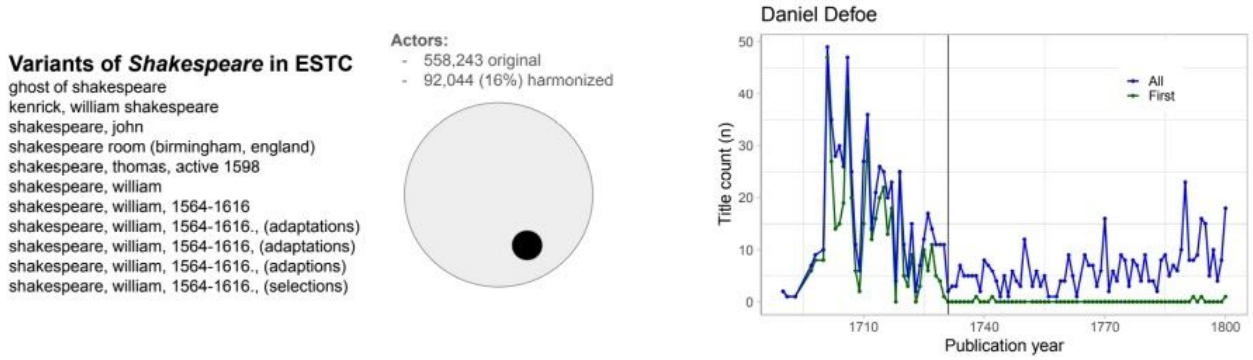

Fig. 1. Impact of data harmonization when using bibliographic records as research data. One of our major data harmonization efforts has been the unification of information of actors in the British printing industry from 1470 to 1800. This means that we have automatically extracted publisher, printer and bookseller information from the imprints and harmonized the data so that the widely scattered actor information is uniformly mapped. This means that we have given each actor a unique identifier, and matched different incidents together. **Left** Shakespeare name variants in the ESTC data before harmonization. **Middle** Drastic effect of the harmonization of all actors, where more than half million scattered entries are reduced to only 16% of the original data due to the linking of the actors together and discarding invalid entries. **Right** Another main effort of ours with respect to ESTC has been to map the information of different titles together so we can study different editions and reprints. We have an example of how this can change the way we study individual authors. Daniel Defoe's (1660-1731) early publishing career is dominated by topical pamphleteering on political matters when the number of all unique published titles (blue) is high, but these consist of mainly first editions (green) where the high frequency of new works indicates topicality. This changes already during Defoe's last years when his literary status becomes more established, especially due to *Robinson Crusoe* (1719), but also *Moll Flanders* (1722), and other literary pieces. What we notice is that in his posthumous publication record the role of a political pamphletist is overtaken by his modern literary image. Also what we witness is some of Defoe's topical poetry turning into literature, for example, *True Born Englishman* (1701) that was printed several times over the eighteenth century.

Further reading:

Leo Lahti, Jani Marjanen, Hege Roivainen & Mikko Tolonen (2019) Bibliographic Data Science and the History of the Book (c. 1500–1800), Cataloging & Classification Quarterly, 57:1, 5-23, DOI: 10.1080/01639374.2018.1543747

Mikko Tolonen, Leo Lahti, Hege Roivainen & Jani Marjanen (2019) A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828, Historical Methods: A Journal of Quantitative and Interdisciplinary History, 52:1, 57-78, DOI: 10.1080/01615440.2018.1526657

Helsinki Computational History Group website:
https://www.helsinki.fi/en/researchgroups/computational-history