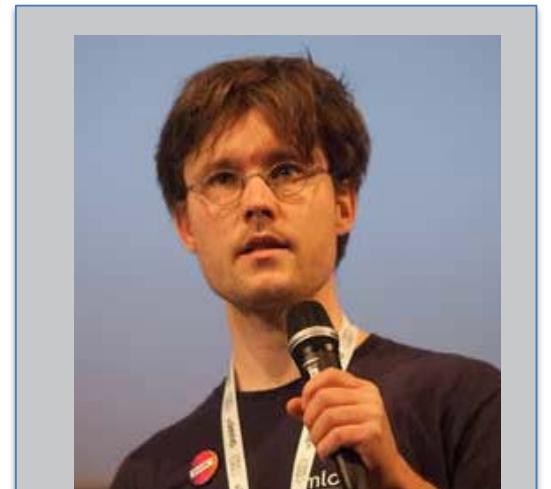# Seminar Announcement

Book History and Bibliographic Data Science

Date: 6 September 2019
Time: 2 p.m.
Venue: Classroom 4, SBS
Hosted By: Asst Prof Jarkko Salojärvi

Speaker:
Assoc. Prof. Leo Lahti
University of Turku

Library catalogues contain rich information on historical patterns in knowledge production, and their research potential has been debated for more than 50 years. Large-scale analysis of these data collections can provide novel ways to investigate classical research hypotheses in intellectual history. A systematic research use of bibliographic metadata has proven to be challenging, however. The lack of scalable solutions to improve and verify data quality, completeness, and representativeness has been a major bottleneck for large-scale analysis. The emerging paradigm of bibliographic data science provides targeted algorithmic methods in this area. We have recently integrated metadata across four large bibliographies and altogether 2.64 million harmonized entries in the period c. 1500–1800. 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. Moreover, a combination of open data and open analysis workflows allows us to demonstrate the opportunities of fully open bibliographic data science based on The Finnish National Bibliography that has been recently licensed in an openly licensed, machine-readable format. We have estimated the long-term development of book formats, which reflects shifts in reading habits and public communication over time. 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. 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 historical research.