Automating data curation for the Finnish national bibliography Fennica

Julia Matveeva
University of Turku, Finland
yulia.matveeva@utu.fi

Osma Suominen
National Library of Finland, Finland
osma.suominen@helsinki.fi

Leo Lahti
University of Turku, Finland
leo.lahti@utu.fi

Abstract

The consortium Digital History for Literature in Finland (Research Council of Finland, 2022–26) differs from earlier research on Finnish literary history by making use of digital collections and new methods in data science, which enable the use of the collection as a whole.1 Here we present a dedicated bibliographic data science framework tailored for the specific context of the consortium research purposes. We will examine how data science methods can improve our understanding of literary history and how it's told, and how reliable the information can be. [1,2,3] The Finnish National Bibliography, Fennica, consists of over one million records from 1488 to the present and includes diverse data types such as books, newspapers, maps, and other documents from 1488 to the present day. The source data contains ambiguous information, missing or erroneous entries, however. Any refinement efforts will include context-specific choices that depend on the research use case. We have previously shown how selected subsets of the collection can be refined automatically to support large-scale statistical analyses of book printing during the years 1500-1800[1, 3]. In our present version of the workflow2, we've scaled up the previous analyses of 70 thousand records to cover all Fennica records and improved the data curation workflows. To cater to the research objectives of the project, we've integrated signum data and created a focused subset covering the years 1809-1917 for each metadata category. Furthermore, we've added a genre subfield derived from a broader leader field to enable genre identification at a bibliographic level, specifically focusing on books. Our aim has been to replicate a curated list, adhering to predefined criteria such as UDC classification, language, genre, and signum data. This automated process mirrors and supports the manual list creation method utilized by the Literary History subproject within the consortium. Future efforts could encompass the integration of further complementary sources, such as the rich information on authors, publishers, and geographic places available in the public domain. These efforts contribute to achieving the consortium's goals, which involve leveraging digital collections and methodologies to broaden the conventional understanding of Finnish literary history. Specifically, we aim to map Finnish and Swedish language fiction from the 19th century into an enriched format conducive to large-scale statistical analyses and the development of reproducible data science workflows.

Keywords: digital humanities, workflow, bibliography, fennica, data

REFERENCES

1. Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly,* 57(1), Special Issue. https://doi.org/10.1080/01639374.2018.1543747

2. Lahti, L., Mäkelä, E., & Tolonen, M. (2020). Quantifying bias and uncertainty in historical data collections with probabilistic programming. *CEUR Workshop Proceedings on Computational Humanities Research, 2723, Short* Paper 46. https://ceur-ws.org/Vol-2723/short46.pdf

3. Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (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. https://doi.org/10.1080/01615440.2018.1513177