Scalable refinement of the Finnish national bibliography for large-scale statistical analysis

Julia Matveeva, Akewak Jeba, Veli-Matti Pynttäri, Kati Launis, Osma Suominen, Leo Lahti

Keywords: bibliographic data science, digital humanities, national bibliographies, open data science, workflow

Statistical analyses of bibliographic metadata catalogs can provide quantitative insights into large scale trends and turning points in publishing patterns, enriching, and even challenging the prevailing views on the history of knowledge production (Lahti, 2019a). The use of bibliographic catalogs has become a well-established tool in literature history and helped to renew research methodology (Umerle, 2023). However, the efficient utilization of large-scale data collections as research material depends critically on our ability to critically evaluate data representativeness, completeness, quality and trustworthiness. Our earlier work has demonstrated how remarkable fractions of the bibliographic metadata curation and analysis process can be automated through dedicated bibliographic data science workflows (Lahti, 2019b, 2015; Tolonen, 2016, 2019).

This study presents further development of an open and scalable data science workflow to support literary research using the Finnish National Bibliography, Fennica. The scalability of the solutions varies by data type, and the refinement process must strike a balance between accuracy and scale. Our reproducible workflows emphasize transparency, consistency, and provenance as key elements of this process; we show how standardized refinement procedures and automated generation of versatile statistical summaries of the refined data can be used to monitor the curation process while supporting in-depth statistical analyses and modeling of publishing patterns over time and geography.

We present good practices and conceptual approaches for bibliographic data refinement and demonstrate how enriched national bibliographies can offer a data-rich perspective on Finland's literary history. This study particularly focuses on Finland's Grand Duchy era (1809–1917) literary analysis, contrasting manual and automated data extraction methods. The dataset, sourced from the National Library of Finland, records in Fennica from 1488 to the present day and numerous fields and subfields. We only approach around 50 fields to cater for our research needs. The workflow employs tailored methods to standardize key metadata fields, including author information, language, publisher, publication place, classification schemes, genre fields such as call number, UDC, control field and index term/genre, title, physical dimensions, and gender. These functions harmonize inconsistencies, remove ambiguities, and integrate supplementary information from external databases, ensuring high data fidelity.

The refined dataset reveals insights into Finnish publishing history, addressing gaps in metadata completeness and quality. Enrichment from external collections and complementary sources,

including Finna, Kanto, and Finto, helps mitigate limitations such as missing author information, ambiguous publisher and publication place data, and the absence of gender classification. Additionally, UDC numbers were converted to words using Finto vocabulary via web scraping.

The results highlight the effectiveness of automated bibliographic data refinement in supporting large-scale research. Key outputs include a comprehensive bibliographic data science workflow, harmonized metadata dataset for research applications, and novel solutions for semi-automatic curation of national bibliographies. Informative data summaries facilitate quality control and bibliographic analysis while enabling focused studies on specific periods. The approach can be adapted for various temporal, geographic, and thematic analyses.

References:

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019a). Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly, 57*(1). Routledge.

Lahti, L., Vaara, V., Marjanen, J., & Tolonen, M. (2019b). Best practices in bibliographic data science. *Proceedings of the Research Data in the Humanities Conference 2019*.

Lahti, L., Ilomäki, N., & Tolonen, M. (2015). A quantitative study of history in the English Short-Title Catalogue (ESTC) 1470–1800. *LIBER Quarterly, 25*(2).

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*. Routledge.

Tolonen, M., Ilomäki, N., Roivainen, H., & Lahti, L. (2016). Printing in a periphery: A quantitative study of Finnish knowledge production, 1640–1828. *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków.

Umerle, T., Colavizza, G., Herden, E., Jagersma, R., Kiraly, P., Koper, B., Lahti, L., Lindemann, D., Łubocki, J. M., Malínek, V., Milanova, A., Péter, R., Rißler-Pipka, N., Romanello, M., Roszkowski, M., Siwecka, D., Tolonen, M., & Vimr, O. (2023). An analysis of the current bibliographical data landscape in the humanities: A case for the joint bibliodata agendas of public stakeholders. *Czech Academy of Sciences*.