Authors: *Julia Matveeva (University of Turku)*, Kati Launis (University of Eastern Finland), Osma Suominen (National Library of Finland), Leo Lahti (University of Turku). Thank you, Akewak Jeba, Teo Dallier, Elliot Gaudron-Parry, Pyry Kantanen (University of Turku). Acknowledgments: The project is supported by Research Council of Finland (DHL-FI 2022-2026 & FIN-CLARIAH FIRI).

Scalable data science workflow for the Finnish national bibliography (Fennica)

Automated workflow

DESIGN:
Semi-automated data refinement and analysis workflow implemented in Quarto and published on CSC Rahti as Gitbook

ANALYSIS:
Combining close and distant reading by literature historians

Fennica metadata conversions: statistical monitoring and analysis
URL: https://fennica-fennica.rahtiapp.fi/

Data

Fennica is a comprehensive database dedicated to the Finnish national imprint. It has records spanning from 1488 to the present. Fennica encompasses various types of materials, including: books (dating back to 1488), newspapers (since 1711),collections, maps, audiovisuals, electronic materials.

The original data is in the standard MARC21 format, and has numerous fields and subfields. Our focus is on fields relevant for Literary History.

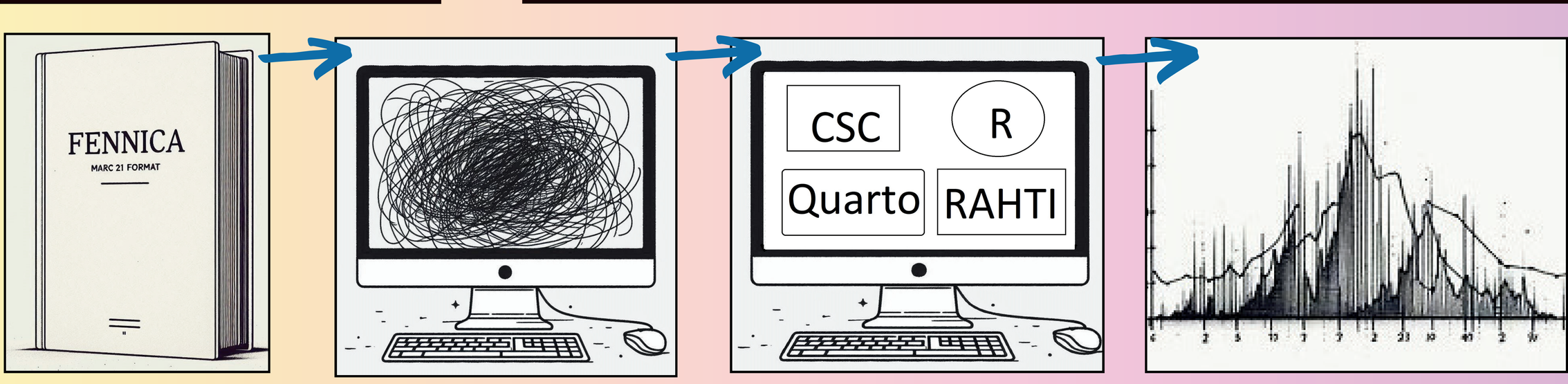
Data science methods

Data Cleaning and Refinement: We cleaned the data to handle missing entries and standardize formats. Refinement strategies were customized for specific research use cases, ensuring data integrity is preserved according to the analytical requirements.

Scalability and Automation: We developed scalable workflows and algorithms capable of efficiently processing over one million records. This ensured consistent and reliable data refinement across large datasets.

Statistical Analysis and Modeling: We employed statistical techniques to identify various trends, including temporal, geographic, and thematic patterns. Automated statistical summaries and visualizations facilitated continuous monitoring and supported comprehensive analysis.

Data Integration: We incorporated Signum data into our dataset to enhance the study of literary history. This integration allowed for deeper insights and enriched our research capabilities.



Harmonized fields

- Author's name, Author's lifetime
- Language
- Publisher, Publication place, Publication time
- Physical dimensions, Physical extent
- Signum, Types of Content, UDCN
- Title, Remainder of Title

Key Statistics

- 1187813 records
- Publication years: 1488-2024
- 13 different types of records
- 196891 unique authors
- 170 unique languages
- 4947 unique publication places
- 156327 unique publishers

Background Literature

Bibliographic Data Sience and the History of the Book (c. 1500–1800)
Lahti L., Marjanen J., Roivainen H., Tolonen M.
Cataloging and Classification Quarterly, 2019