Subset Selection in Bibliographic Research:

Exploring the Boundaries of Automated and Manual Curation

DH2025, Lisbon
18.07.2025

DIGITAL HISTORY FOR LITERATURE IN FINLAND (WWW.DHL-FI.UTU.FI):

JULIA MATVEEVA, UNIVERSITY OF TURKU

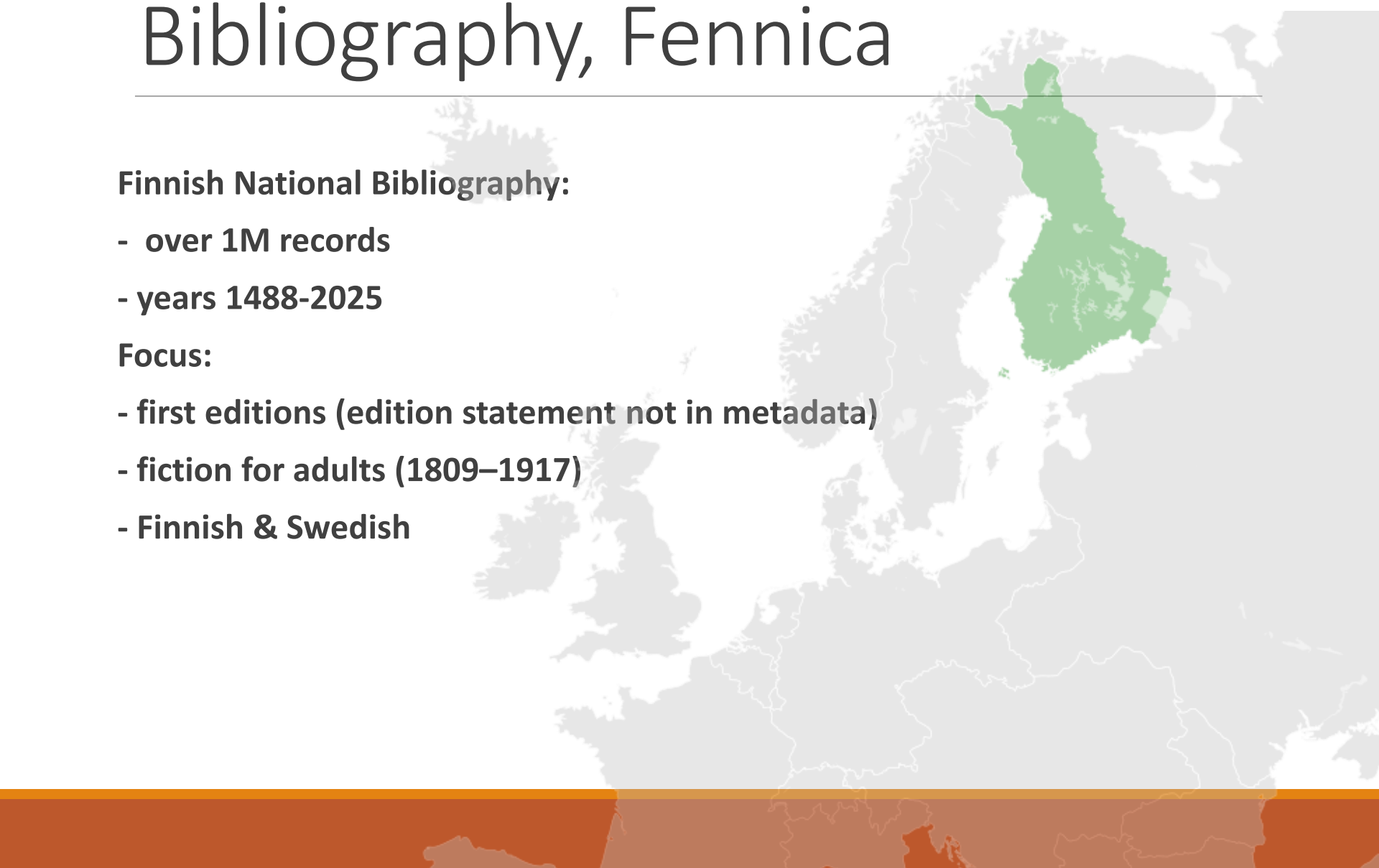
VELI-MATTI PYNTTÄRI, UNIVERSITY OF EASTERN FINLAND

OSMA SUOMINEN, NATIONAL LIBRARY OF FINLAND

KATI LAUNIS, UNIVERSITY OF EASTERN FINLAND

LEO LAHTI, UNIVERSITY OF TURKU

Data Source: Finnish National Bibliography, Fennica



Finnish National Bibliography:

- **over 1M records**
- **years 1488-2025**

Focus:

- **first editions (edition statement not in metadata)**
- **fiction for adults (1809–1917)**
- **Finnish & Swedish**

Manual curation

Primary: Books, Language (Finnish or Swedish), Year (1809-1917), Call number (genre = fiction).

Exclusion and Enrichment: Remove literature for children and translations, integration of records from Reenpää collection

Filtering: first editions

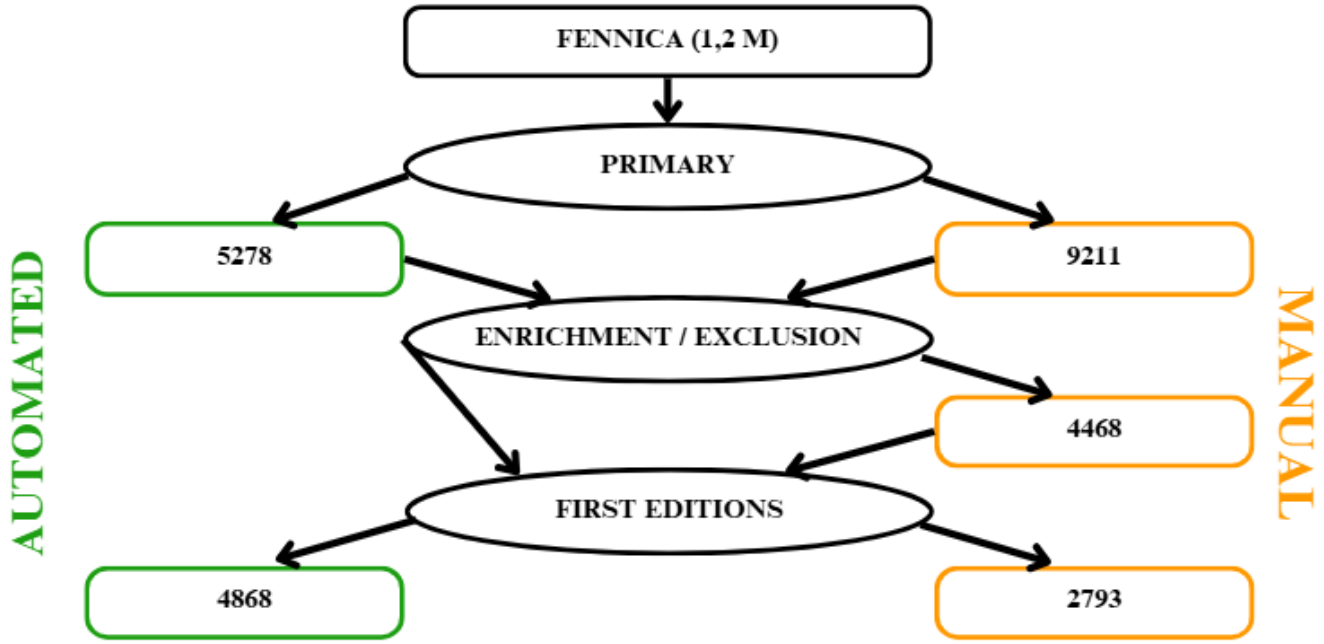
How?

- SQL search in Koha, expert review in Excel
- by hand
- cross-referencing with historiography

Automated curation

- 1,2M dataset from library data dump in MARC21
- Dataset for 1809-1917 is ~70k long
- Basic and custom R packages
- **Primary and Enrichment :** Books, Language (Finnish or Swedish), Year (1809-1917), Fiction (008, Call number, UDC, 655a)
- **Exclusion:** Remove literature for children and translations by pattern and word recognition
- **Heuristic filtering** (first publication of the same combination of "id + author + title (a,b,n) + year")

Subset Selection Flow



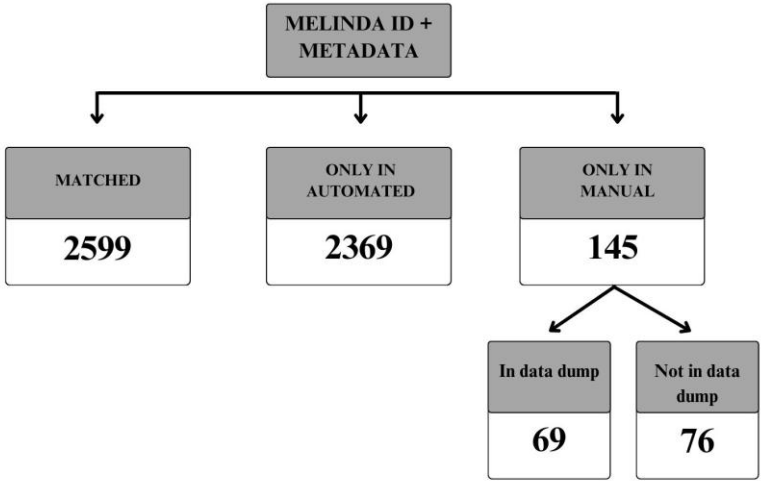
Comparison Results

Exact and pattern matching (thorough harmonization of

"**id** + **author + title (a,b,n) + year**"): removing punctuation and lowercasing

2,599 exact matches (95.3%).

2,369 titles only in the automated list.



145 only in manual:

69 titles rejected by automation

due to metadata issues.

76 not in data dump dataset

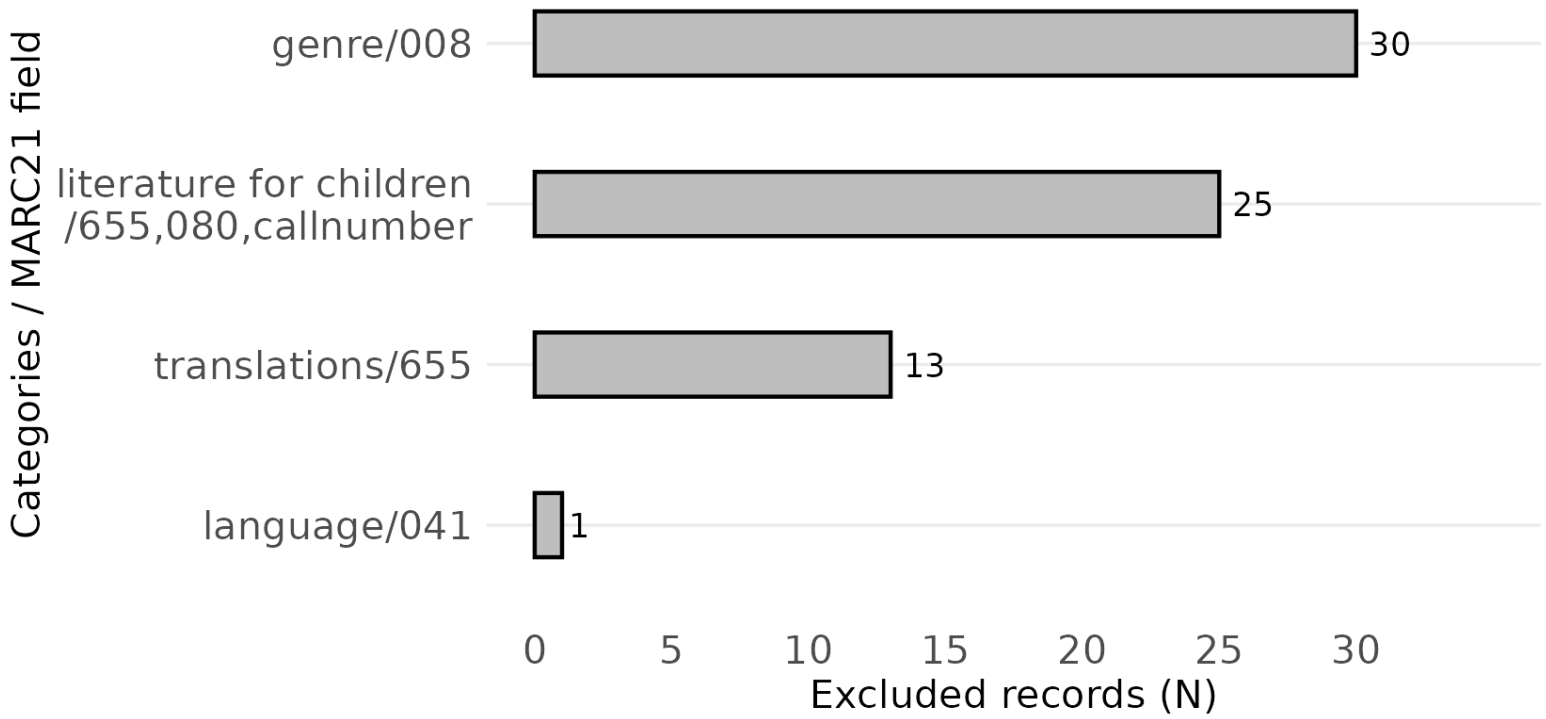
Only in automated: 2369

- 374 can be added to the list of first editions according to experts 🎉
- 1,324 were rejected

The most prevalent reasons for discarding titles from the automated list were **place of publication** (only books published in Finland or, in some cases, in Sweden were included) and **books comprising already published items** (collected works or anthologies).

- 712 were marked as “unsure”

Records discarded from the manual list



Metadata Case Examples

Example 1: Fiction or Non-fiction (Rydman).

Title: **Images** and **Stories** from the Caucasian Steppes and Mountains

Genre/008: Non-fiction

Call number: Finnish language fiction

UDC: Finnish language fiction

655a: Travel literature / Non-fiction

Example 2: Original work flagged as translation (by Aura (Betty Elfving)).

Title: Vuosisatain perinto.

Genre/008: Novel

Language: Finnish

Language_original: NA

Call number: Finnish language fiction

UDC: Finnish language fiction

655a: **Finnish-Swedish literature; Translation**

Strengths & Limitations

Automation: reproducible, scalable, fast (seconds), overfiltering, surfaces overlooked records by manual

Manual: nuanced, culturally informed, laborious, time-consuming, difficult to reproduce

Metadata quality is a major bottleneck.

Conclusion & Takeaways

Hybrid methods produce best results: automation first, manual second

Automation reveals hidden records & gaps in the metadata

Experts oversight remains essential

Metadata quality affects automation

Future Work

Improve criteria: page counts, title fields, publication places

Use NER/ML for genre classification using title fields.

Support metadata quality improvement efforts.

Thank You!

Contact: Julia.Matveeva@utu.fi

Harmonized data: https://fennica-fennica.2.rahtiapp.fi/

UNIVERSITY OF EASTERN FINLAND

1640 KANSALLIS KIRJASTO

CSC

UNIVERSITY OF TURKU

SUOMEN AKATEMIA