What Linked Data can tell about Geographical Trends in Finnish Fiction Literature

– Using the BookSampo Knowledge Graph in Digital Humanities

Telma Peura¹,², Petri Leskinen¹ & Eero Hyvönen¹,²

¹ Semantic Computing Research Group (SeCo), Aalto University, Finland;
² Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland
Introducing BookSampo (1)

Part of the Sampo series
- building a national ontology (FinnONTO) and infrastructure for cultural heritage (CH) data that can be accessed through a semantic portal (Hyvönen, 2020)
- open source
- based on semantic web, residing on a SPARQL endpoint
- unifying and connecting heterogenous data silos of CH data
- Three components of model building
  1) LD creation and publishing model
  2) an interface with multiple perspectives
  3) two-step filtering cycle

Table 1
Sampo Model Principles P1–P6

<table>
<thead>
<tr>
<th></th>
<th>Principle</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>Support collaborative data creation and publishing</td>
</tr>
<tr>
<td>P2</td>
<td>Use a shared open ontology infrastructure</td>
</tr>
<tr>
<td>P3</td>
<td>Make clear distinction between the LOD service and the user interface (UI)</td>
</tr>
<tr>
<td>P4</td>
<td>Provide multiple perspectives to the same data</td>
</tr>
<tr>
<td>P5</td>
<td>Standardize portal usage by a simple filter-analyze two-step cycle</td>
</tr>
<tr>
<td>P6</td>
<td>Support data analysis and knowledge discovery in addition to data exploration</td>
</tr>
</tbody>
</table>

Hyvönen et al. (2022)
What does BookSampo contain?

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary Works</td>
<td>93,000</td>
</tr>
<tr>
<td>Editions</td>
<td>127,000</td>
</tr>
<tr>
<td>Book Covers</td>
<td>27,000</td>
</tr>
<tr>
<td>Fictional Characters</td>
<td>19,000</td>
</tr>
<tr>
<td>Contemporary Reviews</td>
<td>15,000</td>
</tr>
<tr>
<td>Weblinks</td>
<td>10,000</td>
</tr>
<tr>
<td>Literary Series</td>
<td>2,900</td>
</tr>
<tr>
<td>Literary Awards</td>
<td>2,700</td>
</tr>
<tr>
<td>Literary Award Series</td>
<td>200</td>
</tr>
<tr>
<td>Movies</td>
<td>1,100</td>
</tr>
<tr>
<td>People (e.g. Authors)</td>
<td>29,000</td>
</tr>
<tr>
<td>Author’s Pictures</td>
<td>2,600</td>
</tr>
<tr>
<td>Publishers</td>
<td>2,600</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>Nr. of Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Literary Works</td>
<td>207,771</td>
</tr>
<tr>
<td>Editions</td>
<td>213,161</td>
</tr>
<tr>
<td>Edition Parts</td>
<td>79,447</td>
</tr>
<tr>
<td>All Editions</td>
<td>285,518</td>
</tr>
<tr>
<td>Book Covers</td>
<td>112,971</td>
</tr>
<tr>
<td>Main Characters</td>
<td>45,397</td>
</tr>
<tr>
<td>Contemporary Reviews</td>
<td>14,644</td>
</tr>
<tr>
<td>Weblinks</td>
<td>25,017</td>
</tr>
<tr>
<td>Literary Series</td>
<td>8,374</td>
</tr>
<tr>
<td>Literary Awards</td>
<td>6,310</td>
</tr>
<tr>
<td>Literary Award Series</td>
<td>290</td>
</tr>
<tr>
<td>Movies</td>
<td>2,010</td>
</tr>
<tr>
<td>People (e.g. Authors)</td>
<td>62,207</td>
</tr>
<tr>
<td>Author Pictures</td>
<td>4,140</td>
</tr>
<tr>
<td>Publishers</td>
<td>5,455</td>
</tr>
</tbody>
</table>

Mäkelä et al. (2013)

August 2022
Introducing BookSampo (2)

- a collaboration between Finnish public libraries and SW researchers
  - started in 2007
  - published 2011
- original data came from a dump from the Helsinki metropolitan area libraries
- new material automatically converted every night from BTJ Finland Ltd (Mäkelä et al. 2013)
- Kaunokki ontology for fiction literature
- Maintained by a team of librarians
  - improves the quality of the metadata
- contains practically all literature of Finnish public libraries (Mäkelä et al. 2011)
  - a good data source for DH research
Case Study: Geographical diversity in novels

Distant Reading (Moretti, 2013)
- adopting a quantitative perspective to literature
- capturing large patterns in the literary production

Literary Geography (e.g. Evans & Wilkens, 2018)
- in terms of author nationalities and languages
  - extra-textual
- in terms of novel settings
  - contents

How has the geography of novels published in Finnish developed in the past 50 years?
Observations on data quality

Lack of coherence:

- over 600 URIs for genres, and over 400 themes
  - without a clear hierarchy
- “everything that helps in searching”
- e.g. How to group war literature?
  - 2nd WW ⇒ wars (events) ⇒ societal events
  - hostilities (warfare) ⇒ societal events
  ⇒ events
- wartime ⇒ societal periods of time
  ⇒ time periods
- the cold war ⇒ mutual action ⇒ action

Example: detective stories on a map (Hyvönen et al., 2022)
Data preprocessing and methods

SPARQL-querying the BookSampo SW and constructing necessary subgraphs

- constructing a sub-graph with the wanted attributes
  - dealing with annotation errors and irregularities to simplify the query process
  - e.g. Scottish, English, Welsh ⇒ Brits; Ålandians, Swedish-speaking Finns ⇒ Finns
  - Sami people, Finnish Swedes
  - removing duplicated nationalities (several URIs)
  - unification of the language codes
    - most based on lexvo, some irregularities
- Jupyter Notebooks
  - SPARQLWrapper, RDFLib
- Filtering on publication year (1971-2020)
Unbalanced representation of languages and nationalities
Unbalanced representation of languages and nationalities
A Gender Bias in Language Diversity
A similar pattern in nationalities
Geography in the novels written by Finnish authors

Only about ⅓ of yearly publications is annotated for the geographical setting.

Grouping place names by country and year.

Russia the 2nd most popular country in almost every year.

The proportion of Finland slightly decreasing.
Geography in the novels (2010s)
What do authors write about when they write about Russia?

Clustering based on genre and theme keywords

1) biographical, historical war novels, memories  
   48/96/144
2) war literature, societal novels  
   5/110/117
3) personal development, relationships, moving, psychological  
   42/30/72
4) thrillers, crime fiction, espionage  
   6/52/61
5) women, historical romance  
   23/5/28

⇒ Female and male authors have different perspectives on Russia in the context of novels
On the map

Novels about Russia actually tell about Finland.
Focus on Karelia.
Confirms the previous claim (Hallila & Hägg, 2007) that Finns are interested in historical, national topics.

The country of Russia as the annotated place, not a specific location.
Conclusions and Future Directions

The BookSampo Knowledge Graph can be used in DH research, but it requires careful preprocessing and compromising on detailed/specific claims. Novels seem to **become more diverse** both in terms of extra-textual indices and textual settings. Bias in the annotations?

- Geographical setting annotated only in a third of the novels
- **When is the geographical setting seen important?**
  - E.g. genre literature
  - **When should it be annotated?** How fine-grained?

Insights from this exploratory research can help to improve the annotation practices further. Collaboration between DH and SW researchers, and librarians is needed.
References


Thank you! Questions?