Challenges in managing semantic annotations in harvested research objects in a national CRIS context

Tommi Suominen - https://orcid.org/0000-0002-4382-7425, CSC - IT Center for Science
Joonas Kesäniemi - https://orcid.org/0000-0002-3770-0006, CSC - IT Center for Science
Katja Mankinen - https://orcid.org/0000-0001-7357-9648, CSC - IT Center for Science
Contents

• Research.fi

• Semantic annotation

• Word clustering

• Challenges in maintaining annotations, when the background changes
The Finnish Research Information Hub

• The National Research Information Hub project was launched in 2017
• Owned and financed by the Ministry Culture of Finland that has authorised the implementation to CSC – IT Center for Science Ltd.
• Information is publicly and freely available in www.research.fi launched in 2020
Aims of the Finnish Research Information Hub

- To describe the Finnish research ecosystem and act as an information base for science policy
- To support the reuse of existing research information in other services and systems
- To help to find Finnish research and researchers
- To support the merit of researchers and increase the visibility of their research output and activities

R&D analytics
Visualizations
Business Intelligence Landscape

APIs
Shared data model

Research.fi portal
Semantic search
Linked data

Researcher's Profiles
Traditional research outputs
Other research activities
1. The details of researchers, publications, research datasets, research projects and research infrastructures will soon be available in one place.

2. There will be less reporting and administrative work as the data is available in a single location and information flows freely between services.

3. Results of both publicly and privately funded research will be openly accessible.

4. This will benefit researchers, research organizations, funding agencies, public administration and citizens.
Semantic concept annotation and search

Implementation by: Joonas Kesäniemi
Relationships between linkable items (research objects)

• Subject – predicate – object (triplet)

• Depending on the subject and object in question, different predicates will apply. This is fine in the linked data world. In the graph sense, this implies more then two linkable items being simply connected.

---

Fenner, M., Aryani, A., 2019. Introducing the PID Graph — FREYA [WWW Document]. https://doi.org/https://doi.org/10.5438/jwvf-8a66
Also links from research objects to concepts form triplets

- Can also be used in reverse, research objects related to a specific concept:
  - Semantic research object search
  - Navigating research outputs using concept relationships

https://sanastot.suomi.fi/concepts/3bd6c2-e57f-49c1-b04-e37eff042834/concept/8395b72f-7015-4d77-5b0-b-e1699d79c867
Semantic annotation in research.fi

• Semantic annotation is based on utilising a known **ontology** and an **annotator** that has been trained to annotate textual content with concepts.

• These concepts are consequently linked as metadata related to a research object.

• In our context, automated concept annotation uses the **YSO-ontology** using the **Annif-annotator**.

• At first, we focus on annotating the funding decisions, based on their titles and descriptions.
Semantic search based on the annotations

• A search with the concept "forest management" will find a funding decision, with the term "metsänhoito" (Finnish) used in the description.

• This uses the different language labels of a concept that is a direct hit for the searched concept.

• In the next iteration, the search with "forest management" will find the document with the annotated concept “tree planting”, as the two concepts are semantically related.

• Complements the word clustering approach/topic modelling to find funding decisions e.g. by overcoming content language differences.

• First iteration of this implementation is in testing phase

• Next up for development: concept annotation for other research objects
Topic modelling - the word clustering approach
Katja Mankinen, CSC
Topic modelling of research funding decision

- What topics or themes are being researched in Finland or my own organization
- How many projects were funded on climate change/cancer between 2010-2019?
- Which topics/themes received the most funding?

- Unsupervised machine learning methods are used to group text automatically into (often interdisciplinary) topics
- Topic modelling may improve the discoverability of funded research activities (e.g. projects) and offer new insights into the Finnish research field.
- In the Finnish Research Information hub, approx. 30% of the funding decisions are missing information on the field of science that they concern.
- The first iteration of this functionality is in production
Funding decision topic modelling

**Data**
Approx. 7000 decisions (title, description, keywords and field of science)

**Text preparation**
Certain general concepts, frequently occurring non semantic words are removed, language of the description is detected
The text is converted into numerical form (word and document vectors) to facilitate processing

**Topic modelling**
Using and comparing different methods (contextualized topic models, top2vec, BERTopic, hierarchical stochastic block model) to filter potential topics, based on the relationships between words in the descriptions

*Topic* = a cluster of words, which have a semantic relationships of some kind

Based on empirical experimentation, 92 topics were selected, and each funding decision is classified into the most significant topic

Small, large, old, rising, interdisciplinary ... → topics may evolve and new topics may appear
→ the model is updated, when new data is available
Filtering based on identified topic

Projects - 8695

Filter results
Starting year
Organization
Funder
Funding instrument
Field of science

Topic
Keywords
Themes

Academy of Finland's research fields

Identified topic

- Select all
- other
- health care
- manufacturing, production
- education
- sustainable ecosystems, sustainability

Search

Kone Foundation
Salla Tuurinna, University of Turku
2022

Kone Foundation
Jaanke Rekikainen, Tampere University
2022

Academy of Finland
Manohar Kurnar, Aalto University
2022

Academy of Finland
Pelke Vaskema, University of Vaasa
2022

Academy of Finland
Flynn Durabekova, University of Helsinki
2022

Academy of Finland
Benjamin Wilson, Aalto University
2022

Kone Foundation
Tuukka Torpele, Tampere University
2022

Kone Foundation
Kaisra Rask
2022

Kone Foundation
Anni Orola
2022

Maj and Tor Nieselting Foundation
2022

Academy of Finland
Mikael Johansson, CSIC-IT
2022
Visualization

Distribution of projects by year

No search word

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of projects</th>
<th>Amount granted</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2018</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2019</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Legend:
- biomass, bioeconomy, biofuel...
- climate change, adaptation
- education
- energy, electricity, power
- health care
- manufacturing, production
- sustainable ecosystems, sustainable, etc.
Limitations on topic modelling

• All the funding decisions have not been classified to a topic
• All decision are not necessarily tightly connected to the classified topic
• Significant topics may remain undetected/clustered
• Limitations both in the data and in the methods
  o Small and multilingual dataset
  o Very specific topics of research
  o Some decisions have just a title and/or a very short description
  o The topic modelling is based on unsupervised machine learning, small parameter changes may produce different results
• Topic modelling is suitable for mapping research themes on a large scale, but not for determining exact sizes of individual topics.
Challenges in maintaining annotations, when the background changes
Challenges for topic modelling

• As the body of funding decisions defines the clusters, new funding decisions that come in may change the clusters
  o By redefining existing cluster boundaries
  o Introducing a new cluster (and consequently closing another one, as the number of clusters is fixed)

• How to deal with the temporal evolution of clusters? Useful to track historical cluster assignment? For discussion on the “Dynamics of topics” see e.g. Mankinen & Leino 2021
Challenges in annotation management

- Semantic annotation is based on utilizing a known **ontology** and an **annotator** that assigns concepts as metadata to a research object.
- The annotator is trained to annotate specific ontology concepts; but if the ontology changes e.g. through the introduction of new concepts, the annotator is not trained for these new concepts.
- Changes in an ontology must be incorporated into the annotator, to operationalize the newly added concepts.
- New data as such does not impact the ontology or the annotator.
- However, it may be that new content does not fall into the scope of the ontology, and therefore adequate annotation of content with concepts might not be achieved.
The separation of facts and inferred information

• If we have utilized reasoning/inference to create new machine-generated metadata, then if the facts based on which the reasoning was performed, change, the previous conclusions may become invalid. The process of inference needs to be reconducted and the conclusions redrawn.

• This then requires the separation of facts and inferred information, and potentially the linkage of the facts to their conclusions in a traceable manner. (the provenance of the conclusions)

• This would enable re-inference, when we notice one of the dependent facts changed.

• If we are not able to do this, then the recourse is to re-reason the entire body of inferred conclusions
Thank you for your attention!

I am Tommi Suominen. I work at CSC as an information architect. www.linkedin.com/in/tommi-suominen-935b5a8a

I will be coordinating the FAIRCORE4EOSC Horizon Europe project starting in June 2022. I hope to be talking about this at the next euroCRIS meeting.

facebook.com/CSCfi
twitter.com/CSCfi
youtube.com/CSCfi
linkedin.com/company/csc--it-center-for-science
github.com/CSCfi