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

Abstract

Harvested metadata on research objects can include links between the primary domain objects such as organizational identifiers associated with dataset, persons identified with ORCIDs linked to publications and publications connected through ISSNs to publishing channels. This kind of linkage is the bread-and-butter of the CRIS systems and usually comprehensively maintained. When it comes to the more subjective description of a domain object, such as keywords, themes, or subject headings, the issues related to data management and modeling become prominent with challenges such as flexibility of free text keywords as opposed to authoritative, but rigid classification systems. Many CRIS objects also already contain an extensive description of the content, just meant for human consumption, in the form of an abstract or similar summary text. With the help of automated data mining and annotation tools, these textual representations can be processed into structured data. This paper presents the processing pipelines implemented as part of the research.fi portal for automatic linking of different research inputs based on automatically extracted ontology concepts and discusses the implications of utilizing them as part of the research.fi platform. But more than simply discussing the annotation of research objects and the creation of word clusters for representation of the semantic content of research objects, we also discuss challenges related to maintaining the automatically produced metadata, as the utilized ontologies evolve, annotation algorithms develop, connections between research objects and mined word clusters change over time.

Keywords: CRIS, ontology, annotation, linked data

2- page abstract body

Research.fi is a national aggregative CRIS that aims to collect and connect publications, with related funding decisions, research infrastructures and datasets related to each other in some way. The portal content can be browsed from different viewpoints, and the linked data navigated through interconnected links.The Research.fi portal currently connects different types of research outputs using PIDs, when they are directly related to each other in some way. From a service design perspective, the central node in this graph is the researcher. One of the key use cases for the portal design was finding people with expertise in a particular research topic. For challenges such as this, the use of more descriptive keywords, or ontological concepts need to be adopted. Such a search needs to be able to navigate between the different natural languages, and the networks of semantic relationships between concepts.

With Research.fi we have started simple. Many of our funding decisions lack fields of science and keywords. However, nearly all of them have abstracts intended for the human reader. So we have first applied AI-based clustering algorithms to the titles and abstracts of the funding decisions (~16 000 objects) to form thematic clusters of topics based on the frequency of co-occurrence of words found in the texts. Next, we have utilized the Annif annotation tool to extract descriptive ontological concepts of these funding decisions, to eventually allow for implementing a concept based search. In the short run we will be able to implement multilingual searches, and utilize simple related-concept relationships.A roadmap for applying the described automated annotation process across all different research objects types handled by the research.fi portal is being developed. For example, a comprehensive set of publication abstracts have not up to now been freely available due to copyright reasons. A change to this practice has been initiated, to enable collecting also the abstract data in the process of publication data collection.

We are currently using the freely available shared Annif service provided by the National Library of Finland. However, we are actively looking into a possibility to set up our internal instance of Annif as well as setting up our own pipelines for training the annotation models. This would allow us to fine-tune the models to be better suited for our research.fi use case.

As this body of knowledge on concepts describing the research objects grows, we will be able to create a network of research objects parallel to the network formed by research objects' PID-to-PID relationships. This new network is formed by the research objects semantically related to each other, irrespective of whether they were created in studies completely separate from each other, from a people or organizational perspective.

Descriptive keywords or terms can be used to categorize and summarize the research object and they can be very effective in different kinds of search scenarios. However, their effectiveness is hindered by low quality and sparseness of data (also a problem for training AI algorithms). Useful concept annotations are not trivial to come by, and distributing annotation tasks to researchers makes it challenging to maintain consistent content and quality.

If we have utilised reasoning to create machine-generated metadata, then if the facts based on which the reasoning was performed change then it goes to reason that the conclusions would also need to be redrawn. This then requires the separation of facts and reasoned information, and potentially linkage of the facts to their conclusions in a traceable manner.

As we harvest research objects from various sources, we may also receive content annotated using different ontologies then we ourselves use. In a good case, we can utilize an existing and available ontology mapping to bridge different ontologies onto each other. If such are not available, we might be required to create these mappings ourselves. If maintaining the automatically created ontology mappings was a challenge, then a further challenge is most definitely the need to maintain these ontology mappings, and any consequent inferred information that has been produced using them.