Cohesive View to Heterogenous Data through Controlled Vocabularies

Matias Frosterus

Metadata has a key role in making information findable. Well-done descriptive metadata using well-curated controlled vocabularies can help cross barriers of domain, of language, and of time. As an example, places can have different names in different languages and the name might change over time. A document using the Finnish name of a given place at a given time can be automatically linked to a different document annotated using the Swedish name of the same place at a different time - as long as the vocabulary used for the annotation contains the names in both languages and at both times. Using this connection, the researcher interested in the documents concerning the place can find them using a simple query as long as the search engine makes use of the possibilities afforded by the rich vocabularies.

The National Library of Finland, in a project funded by the Ministry of Finance and the Ministry of Culture and Education, has built the national thesaurus and ontology service Finto (http://finto.fi), which provides a centralized service for publishing and utilizing thesauri, light-weight ontologies, and classifications for free. Finto is a direct continuation of the work done on the ONKI system in the FinnONTO research project in Aalto University and the University of Helsinki.

Finto works on top of an open source SKOS browser called Skosmos which is being developed in the National Library. Skosmos allows anyone to set up a service similar to Finto providing the browsing interface as well as a modern REST API which allows the integration of the vocabularies into other systems.

Aside from the service itself, National Library is also continuing the work that FinnONTO research project began in co-ordinating the development of a national cloud of interlinked light-weight ontologies of various domains. At the heart of this construction lies the General Finnish Ontology YSO which contains some 29,000 concepts as of this writing. YSO has labels available in Finnish, Swedish, and English and also contains links to other vocabularies, both Finnish and international.

Expanding on YSO are a number of domain ontologies contributing concepts with a narrower focus and maintained by domain experts in various organizations. Since every domain ontology expands on the upper hierarchy and the general concepts provided by YSO, all the domains are interlinked and relate to one another. Most of the domain ontologies have been built on top of existing thesauri and retain the links to the documents that have been annotated using the original thesauri.

Taken as a whole, this set of interlinked light-weight ontologies, built using common conventions and expressed in a mutually compatible way provides a massive trove of possibilities for making new connections in all sorts of data.

Additionally, Finto hosts various classifications, thesauri, and other vocabularies such as the spatial ontology YSO-paikat. A growing ecosystem of various controlled vocabularies allows for the creation and utilization of descriptive metadata for various aspects of digital humanities research. Making the data of libraries, museums, and archives more interoperable is of obvious benefit but furthermore the vision is to extend this to, e.g., public administration and the legal domain offering an ever richer tapestry of data to discover and explore.