Linking the STCN and Performing Big Data Queries in the Humanities

Wouter Beek, Rinke Hoekstra, Fernie Maas, Albert Meroño Peñuela and Inger Leemans

The Short Title Catalogue, Netherlands (STCN) is a retrospective bibliography containing descriptions of over 200.000 Dutch publications from the period 1540-1800. The STCN is compiled by the National Library of the Netherlands.

In order to facilitate more complex querying capabilities than is possible via the Picarta search interface, we have converted the STCN dataset to Linked Open Data (LOD). Here we report on how we were able to enrich the STCN dataset by linking it to other datasets.

Semantic Web techniques for Digital Humanities

The Semantic Web (SW) approach and the Linked Open Data (LOD) paradigm provide a way for existing structured datasets to be interlinked. It has been argued that the SW approach has specific advantages in the Digital Humanities (DH) domain. Since existing datasets in the humanities are fragmented over a large number of institutions and are often stored in idiosyncratic ways, the combined DH datastore is very heterogeneous. Traditional database techniques require data to be regimented into a uniform format before data is to be meaningfully integrated. The LD paradigm, however, was designed for linking between heterogeneous sources, relating without integrating them.

In converting the STCN to LOD we had the following goals:

- Extend the STCN with relevant related information from other datasets.
- Standardize the vocabulary that is used by the STCN in order to improve the quality and interoperability with other datasets.
- Allow queries to be answered that could not be asked before.

\end{itemize}

We have linked the STCN to the lists of prohibited books by Knuttel and Weekhout, thereby allowing queries to be performed over these heterogeneous data sources. We have also used alignment tools to create fully automated links between authors and their pseudonyms. For the latter alignment tasks we have found 445 such pseudonyms by relating the STCN authors to DBpedia entries.

Since LOD has an explicit schema and a formally defined, model-theoretic semantics, it is possible to perform more complex queries over the LOD-version of the STCN. For this we have written a query layer called humR in SWI-Prolog that uses R in order to perform statistics processing on the data in terms of the domain vocabulary. It is possible to use domain vocabulary terms in formulating the query. For example, one can ask how a certain genre's popularity changes over time. Complex query results can be visualized using standard data visualization tools.

Our research shows that the use of LOD technologies has the potential to transform both the quantitative and qualitative aspects of querying a DH dataset.