SMS: a platform for linking and enriching data for science and innovation studies

In order to improve the data infrastructure for Science, Technology and Innovation Studies, we developed the SMS platform for integrating and enriching data from several heterogeneous data sources. This enables studies to be at the same time large scale (= many entities) and rich (= many variables). The architecture is represented in the figure below (right side). The SMS platform uses Semantic Web and Linked Data technologies. The source data can be confidential or proprietary, data with specific access rules, but SMS also integrates available open data. The main philosophy is that by linking data three things can simultaneously be realized: datasets are (i) combined, (ii) enriched and (iii) entities (people, organizations) can be disambiguated. The data store contains currently some 38 data sets, with research performing organizations, researchers, research projects, patents, publications, but also datasets with geo-boundaries, and with geostatistical data (figure below right side).

Through linking one can create research data about e.g., research organizations from several databases, such as organization characteristics from the ETER database on Higher Education institutions, and performance data from the Leiden Ranking or from Web of Science, and project participations from the Cordis database. Enriching is done e.g., through linking with open datasets such as DBpedia, GeoNames or OpenStreetMap. Disambiguating of organization names can be done by linking several databases, and matching organization names through equivalence of properties (Cordis, Grid, Orgref, Fundref, etc.). Linked data are stored in the SMS data store.

A series of services is implemented on top of the data store: (i) Basic dataset linking; (ii) Homogenizing categories such as research field descriptors; (iii) Geo-localizing to find coordinates for addresses; (iv) Named entity recognition, in order to annotate textual data using knowledge taxonomies; (v) Query services, to retrieve data from different datasets; (vi) Metadata, controlled access, and other services to regulate access.

On top of these services, several user applications exist. The geo-localization application enables the user to enter a list of addresses, and the application gives back geo-coordinates, and geo-boundaries at several levels (figure below – left). As many geo-statistical data are openly available, the system allows the user to (i) add these geo-statistical data as properties of entities (e.g. organizations) within that geographical area, and (ii) define the geo-boundaries in terms of those geo-statistical properties that the researcher finds relevant. This allows the researcher to use other geo-boundaries than the traditional administrative boundaries, or the more
recent introduced Functional Urban Areas by OECD and Eurostat. An example of alternative boundaries is in figure below (right). In the presentation at the conference we will show how this leads to new insights. The selection of the geographical boundaries does in fact influence the geography of innovation one observes.

A second application is the faceted browser (figure below – left side), enabling to browse the linked data for qualitative analysis. In this example, we tried to find structural change in national research systems, characterized by periods of the foundation of many new organizations. Other applications are for link correction, for annotating textual data (figure below right), or for querying data store to build datasets for statistical analysis and visualization.

This extended abstract shows a few of the characteristics of the platform. It enables new ways of data integration, enrichment and analysis. At the conference, we will briefly describe the platform and show one or two examples of how the platform can be used in research projects.

The beta version of the system is now operational. We welcome researchers to visit us and use the platform. Researchers from EU countries can be funded travel and subsistence through the project.