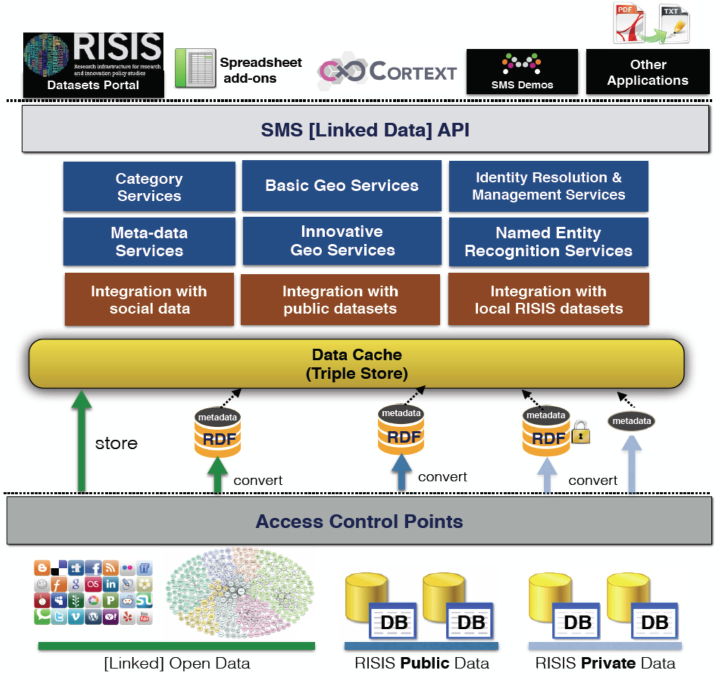
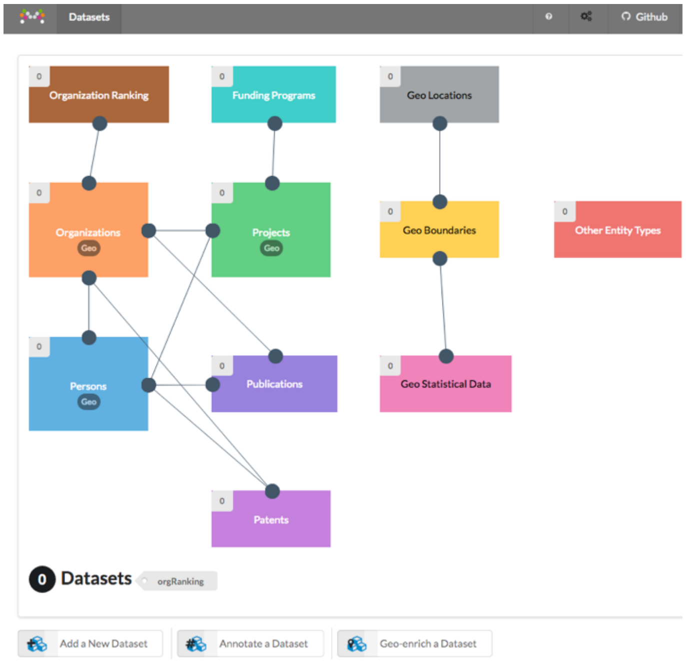
SMS: a platform for linking and enriching data for science and innovation studies

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).

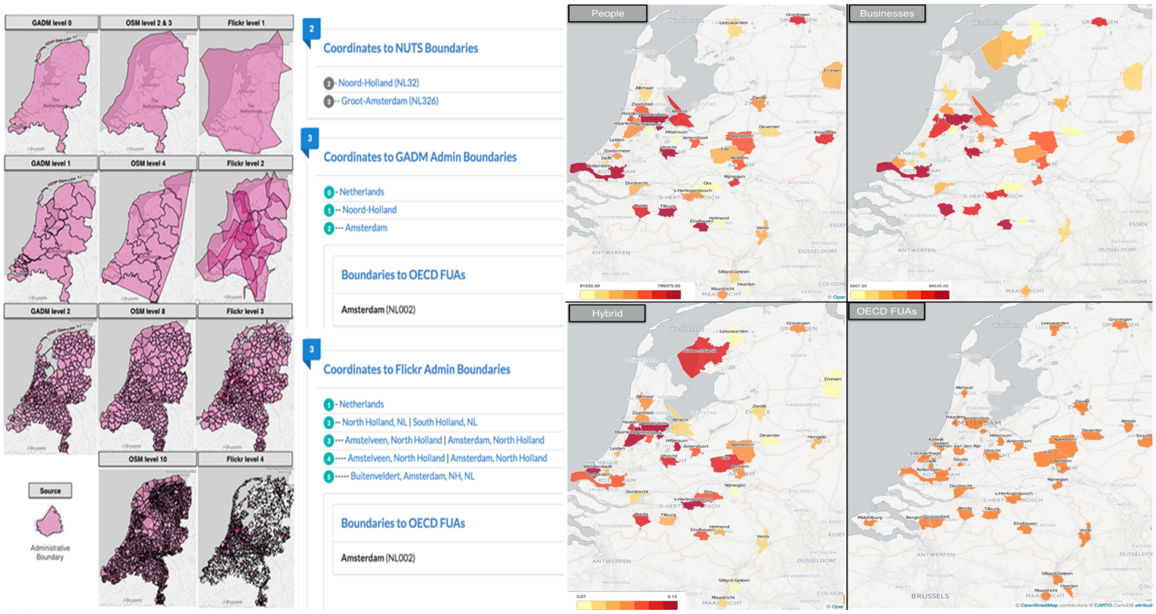




Through linking one can create research data data about e.g., research organizations from several databases, such as organization characteristics from the *ETER* database on Higher Education institutions, and performance data from the *Leiden Ranking* or from *Web of Science*, and project participations 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.



2 Coordinates to NUTS Boundaries

- 2 - Noord-Holland (NL32)
- 3 -- Groot-Amsterdam (NL326)

3 Coordinates to GADM Admin Boundaries

- 0 - Netherlands
- 1 -- Noord-Holland
- 2 --- Amsterdam

Boundaries to OECD FUAs

Amsterdam (NL002)

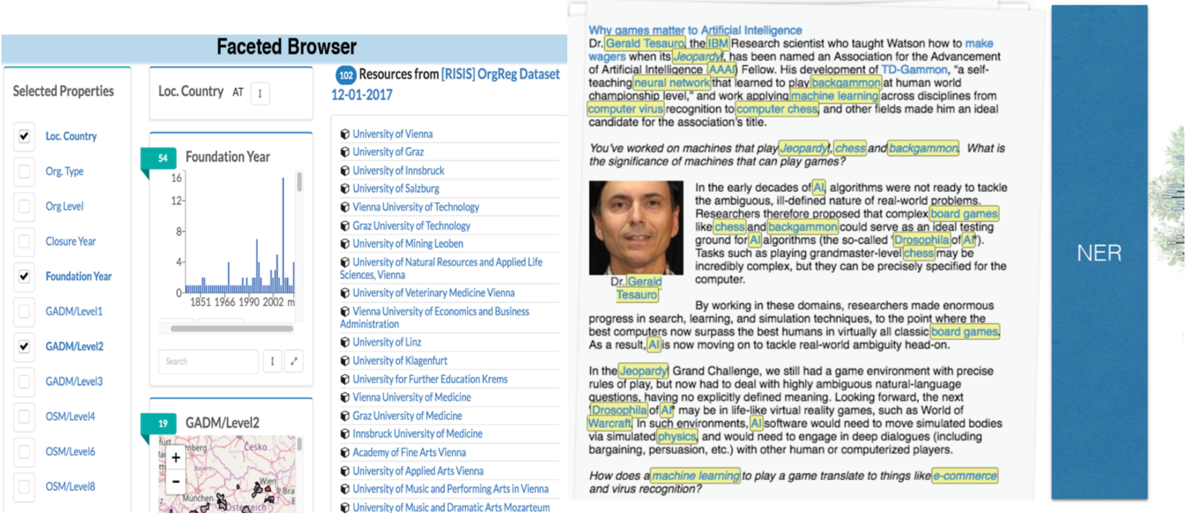
3 Coordinates to Flickr Admin Boundaries

- 1 - Netherlands
- 2 -- North Holland, NL | South Holland, NL
- 3 --- Amstelveen, North Holland | Amsterdam, North Holland
- 4 ---- Amstelveen, North Holland | Amsterdam, North Holland
- 5 ----- Buitenveldert, Amsterdam, NH, NL

Boundaries to OECD FUAs

Amsterdam (NL002)

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.



☑ Loc. Country
□ Org. Type
□ Org Level
□ Closure Year
☑ Foundation Year
□ GADM/Level1
☑ GADM/Level2
□ GADM/Level3
□ OSM/Level4
□ OSM/Level6
□ OSM/Level8

Loc. Country AT

102 Resources from [RISIS] OrgReg Dataset 12-01-2017

- University of Vienna
- University of Graz
- University of Innsbruck
- University of Salzburg
- Vienna University of Technology
- Graz University of Technology
- University of Mining Leoben
- University of Natural Resources and Applied Life Sciences, Vienna
- University of Veterinary Medicine Vienna
- Vienna University of Economics and Business Administration
- University of Linz
- University of Klagenfurt
- University for Further Education Krems
- Vienna University of Medicine
- Graz University of Medicine
- Innsbruck University of Medicine
- Academy of Fine Arts Vienna
- University of Applied Arts Vienna
- University of Music and Performing Arts in Vienna
- University of Music and Dramatic Arts Mozarteum

Why games matter to Artificial Intelligence

Dr. Gerald Tesauro, the IBM Research scientist who taught Watson how to make wagers when its *Jeopardy!*, has been named an Association for the Advancement of Artificial Intelligence (AAAI) Fellow. His development of TD-Gammon, "a self-teaching neural network that learned to play backgammon at human world championship level," and work applying machine learning across disciplines from computer virus recognition to computer chess, and other fields made him an ideal candidate for the association's title.

You've worked on machines that play Jeopardy!, chess and backgammon. What is the significance of machines that can play games?

Dr. Gerald Tesauro

In the early decades of AI, algorithms were not ready to tackle the ambiguous, ill-defined nature of real-world problems. Researchers therefore proposed that complex board games like chess and backgammon could serve as an ideal testing ground for AI algorithms (the so-called 'Drosophila of AI'). Tasks such as playing grandmaster-level chess may be incredibly complex, but they can be precisely specified for the computer.

By working in these domains, researchers made enormous progress in search, learning, and simulation techniques, to the point where the best computers now surpass the best humans in virtually all classic board games. As a result, AI is now moving on to tackle real-world ambiguity head-on.

In the Jeopardy! Grand Challenge, we still had a game environment with precise rules of play, but now had to deal with highly ambiguous natural-language questions, having no explicitly defined meaning. Looking forward, the next 'Drosophila of AI' may be in life-like virtual reality games, such as World of Warcraft. In such environments, AI software would need to move simulated bodies via simulated physics, and would need to engage in deep dialogues (including bargaining, persuasion, etc.) with other human or computerized players.

How does a machine learning to play a game translate to things like e-commerce and virus recognition?

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.