Semantically Mapping Science (SMS) Platform

Ali Khalili[1], Peter van den Besselaar[2], Al Koudous Idrissou[1],
Klaas Andries de Graaf[1] and Frank van Harmelen[1]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, NL
{a.khalili,o.a.k.idrissou,ka.de.graaf,frank.van.harmelen}@vu.nl
[2] Department of Organization Sciences, Vrije Universiteit Amsterdm, NL
p.a.a.vanden.besselaar@vu.nl

Abstract. Up to now, STI (Science, Technology, Innovation) studies are either rich but small scale (qualitative case studies) or large scale and under-complex – because they generally use only a single dataset like Patstat, Scopus, WoS (Web of Science), OECD STI indicators, etc., and therefore deploying only a few variables – determined by the data available. However, progress in the STI research field (and the social sciences in general) depends in our view on the ability to do large-scale studies with often many variables specified by relevant theories. There is a need for studies which are at the same time big and rich. The aim of the Semantically Mapping Science (SMS) platform is to enable enriching and integration of heterogeneous data, ranging from tabular statistical data to unstructured data found on the Web, in order to exploit the huge amount of data that are 'out there' in an innovative and meaningful way.

1 Introduction

Social phenomena generally are complex, and understanding those phenomena requires integrating and analyzing data from multiple sources. Up to now, STI (Science, Technology, Innovation) studies are either rich but small scale (qualitative case studies) or large scale and under-complex – because they generally use only a single dataset like Patstat, Scopus, WoS (Web of Science), OECD STI indicators, etc., and therefore deploying only a few variables – determined by the data available. However, progress in the STI research field (and the social sciences in general) depends in our view on the ability to do large-scale studies with often many variables specified by relevant theories. There is a need for studies which are at the same time big and rich.

In this paper, we present the Semantically Mapping Science (SMS) platform as a means to enable enriching and integrating heterogeneous public and private data, ranging from tabular statistical data to unstructured data found on the Web, in an innovative and meaningful way. SMS is built as an open source platform[3] and is available online at http://sms.risis.eu.

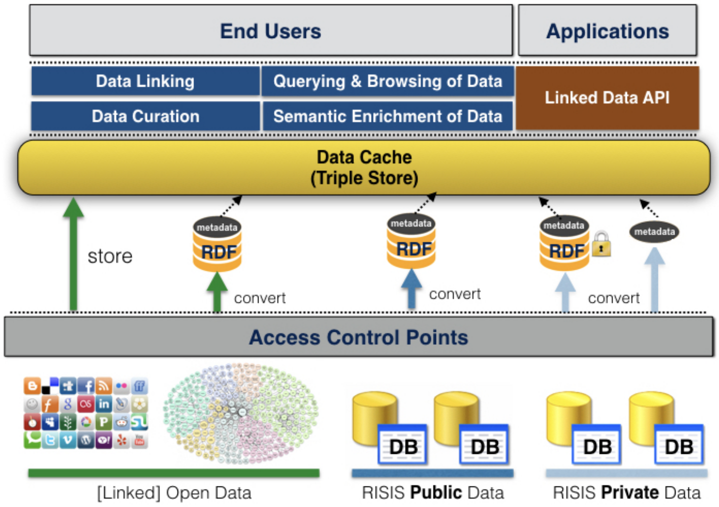


Fig. 1: The SMS Platform Architecture.

2 Architecture

As shown in Figure 1, the SMS platform consists of three main layers: data layer, services layer and application layer. Data layer deals with data conversion, storage and access plans. Service layer provides a set of Web services on top of the created Linked Data to allow developing innovative applications. The application layer is the terminal for end-users who interact with the SMS platform. In this system paper, we briefly describe the main services and applications provided by the SMS platform:

2.1 Conceptual Model

SMS platform at its conceptual model employs an entity-centric approach to interlink heterogeneous datasets in the STI domain. As shown in Figure 2, the following entity types are extracted after analysis of existing RISIS datasets and their related open datasets: Funding Programs, Projects, Publications, Patents, Persons, Organizations, Organization Rankings, Geo locations, Geo boundaries and Geo statistical data. It is also possible to add new entity types based on the research questions which need to be answered by the SMS infrastructure. The main idea is creating a data network by linking and enriching the data, a network which the social science user can access through the faceted browser. By selecting the required entities and properties from the data network, the user gets an overview of the data he/she is interested in. The platform produces in the background the required SPARQL queries to retrieve the selected data from multiple datasets in a required format for further analysis.

[3] https://github.com/risis-eu/sms-platform

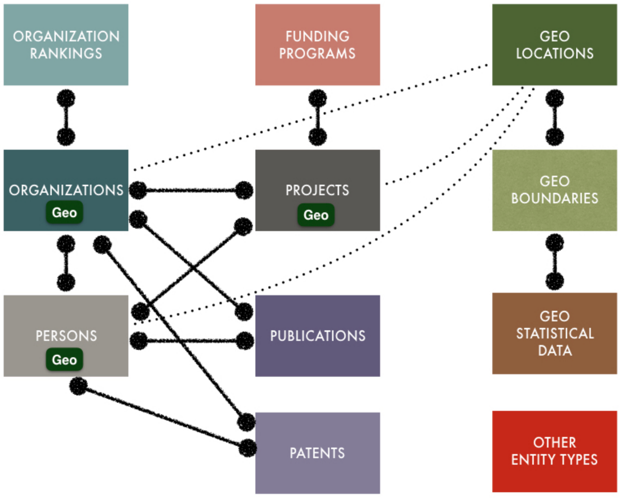


Fig. 2: The Main Entity Types Involved in the SMS System.

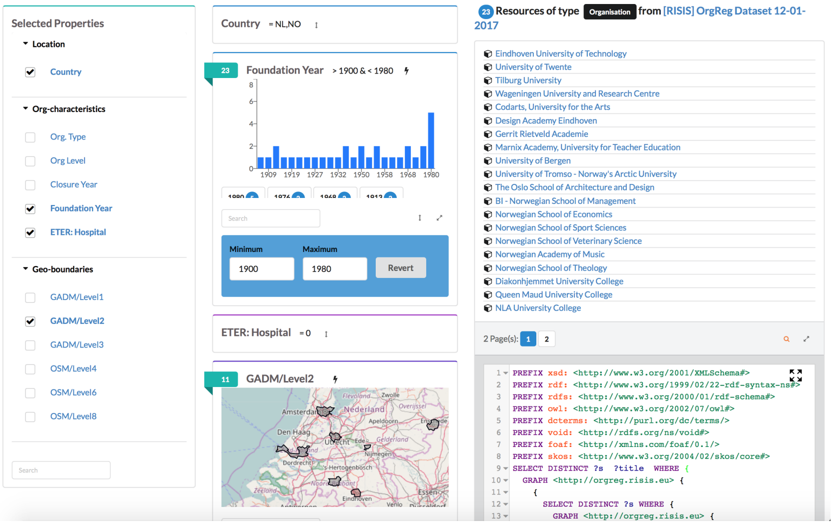
2.2 Data Curation

Metadata helps potential users of a dataset to decide whether the dataset is appropriate for their purposes or not. SMS platform has a collection of various heterogeneous datasets that are not always publicly accessible due to privacy issues, and often require a researcher to be physically at the dataset location. To access these datasets, one needs to be granted an access request. This administrative detour that a researcher has to endure prior to detecting which dataset to use for a particular research question can reduce the number of SMS datasets visitors. It has been shown that research publications that provide access to their base data yield consistently higher citation rates than those that do not. Therefore, to attract more users, to visit and cite RISIS datasets, SMS provides a dataset metadata service and application - modeled using the Resource Description Framework (RDF) - that allows researchers to search for data, and have an in-depth understanding of the data without the need to directly access it [2]. Metadata service powered by an intuitive UI allows dataset holders to describe their datasets in a detailed, consistent and uniform way, store the description and if needed modify the stored metadata. [4] The curated metadata are then reflected on RISIS dataset's portal available at `http://datasets.risis.eu`.

2.3 Browsing and Querying Datasets

One of the objectives in developing the SMS platform was to enable non-Linked Data experts to query and browse RDF datasets without having the knowledge of SPARQL query language. There are currently two main approaches to make information retrieval from SPARQL endpoints more usable: user interaction and natural language (NL). In the category of user interaction-based query generation, faceted browsing user interfaces are well-known techniques which provide

[4] see an screencast of the SMS metadata editor at `https://youtu.be/p_2D3ydcx1U`



Selected Properties

Location

☑ Country

Org-characteristics

☐ Org. Type
☐ Org Level
☐ Closure Year
☑ Foundation Year
☑ ETER: Hospital

Geo-boundaries

☐ GADM/Level1
☑ GADM/Level2
☐ GADM/Level3
☐ OSM/Level4
☐ OSM/Level6
☐ OSM/Level8

Search

Country = NL,NO

23 Foundation Year > 1900 & < 1980

Search

Minimum 1900 Maximum 1980 Revert

ETER: Hospital = 0

11 GADM/Level2

23 Resources of type Organisation from [RISIS] OrgReg Dataset 12-01-2017

- Eindhoven University of Technology
- University of Twente
- Tilburg University
- Wageningen University and Research Centre
- Codarts, University for the Arts
- Design Academy Eindhoven
- Gerrit Rietveld Academie
- Marnix Academy, University for Teacher Education
- University of Bergen
- University of Tromso - Norway's Arctic University
- The Oslo School of Architecture and Design
- BI - Norwegian School of Management
- Norwegian School of Economics
- Norwegian School of Sport Sciences
- Norwegian School of Veterinary Science
- Norwegian Academy of Music
- Norwegian School of Theology
- Diakonhjemmet University College
- Queen Maud University College
- NLA University College

2 Page(s): 1 2

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX void: <http://rdfs.org/ns/void#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT DISTINCT ?s  ?title  WHERE {
  GRAPH <http://orgreg.risis.eu> {
    {
      SELECT DISTINCT ?s WHERE {
        GRAPH <http://orgreg.risis.eu> {
```

Fig. 3: An screenshot of the SMS faceted browser.

a convenient and user-friendly way to navigate through a wide range of data collections [1]. Faceted browsing UIs allow users to find information without a-priori knowledge of its schema [4]. A faceted interface has several advantages over keyword search or NL queries: it allows exploration of an unknown dataset since the system suggests restriction values at each step; it is a visual interface, removing the need to write explicit queries; and it prevents dead-end queries, by only offering restriction values that do not lead to empty results [4].

SMS provides an adaptive component-based faceted browser environment[5] on top of the LD-R framework [3] to allow end-users explore STI related datasets in an integrated way and to incorporate additional features for serendipitous knowledge discovery (see Figure 3 for a screenshot).

2.4 Semantic Enrichment of Data

SMS provides a set of services and applications that allow users to enrich their data by adding complementary data to their current data. There are three categories of data-enrichment services provided:

Named Entity Recognition. Named-entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Given a dataset which has one or more attributes with textual values, SMS NER service can extract named entities from the text and more importantly connect the extracted entities to a knowledge graph or taxonomy (which can then provide more data about those entities). By default, SMS employs DBpedia Spotlight service for NER. However,

[5] see an screencast of the SMS faceted browser at https://youtu.be/9TMLKdGZExY

any arbitrary NER service can be plugged into SMS NER service as long as the output of service is reconciled to SMS named entities annotation model. SMS provides an interactive UI to annotate a dataset using the NER service[6].

Geo-enrichment. Geo-enrichment is an instrument to enrich data by linking through geo-location. Many (open) datasets provide variables that are measured at some level of geographical aggregation: e.g., environmental data, educational data, or socio-economic data. In order to exploit these linking and enriching possibilities, the SMS platform provides a variety of geo-services. The geo-services are based on a series of open geo-resources, such as GADM, OpenStreetMap and Flickr geotagged data. By integrating these geo-resources, the service can give for an entity's address the geo-location up to 11 different levels. One practical application we built for batch processing of addresses is a Google spreadsheet add-on[7] which chains Google Geocoding API with our geo-boundary services. Given addresses in a spreadsheet are enriched with different levels of administrative boundaries and FUAs. The users are then able to export the extracted boundaries and process them in geodata analysis tools such as CartoDB.[8] We have also developed a user interface for automatic geo-enrichment of linked datasets in the SMS platform. The interface allows users to select an existing dataset and geocode the whole dataset by selecting the right attributes in the dataset[9].

2.5 Data Linking

Linking between entities in different datasets is a crucial element of the SMS platform. Whether or not two entities should be considered equal depends not only on their intrinsic properties, but also on the purpose or task for which the entities are used. As an example, to study the success of scientific organizations, STI researchers need to align research organizations across datasets such as GRID[10] and OrgRef[11] that describe organisations across various countries including public and private research organisations. The 3M corporation, a large multinational organisation with a substantial patent portfolio, occurs in both datasets. GRID distinguishes between national 3M branches across six countries *3M (Canada), 3M (France), 3M (Germany), 3M (Israel), 3M(United Kingdom)* and *3M(United States)*, while OrgRef only refers to a single *3M* entity. Should these entities be designated as "the same" across these datasets? It depends. For a study that aims to compare organizations at a global level, all branches of '3M' should be considered the same. Whereas, for a study that compares organizations for a comparison across countries, the Canadian and U.S. branches of '3M' should be considered separately.

SMS provides a novel approach called "Lenticular Lens" for building context-specific links between entities of interest. These links are decorated with rich metadata describing how, why, when and by whom they were generated. As

[6] see an screencast of the NER UI at https://youtu.be/OcYNpVRP9_Q

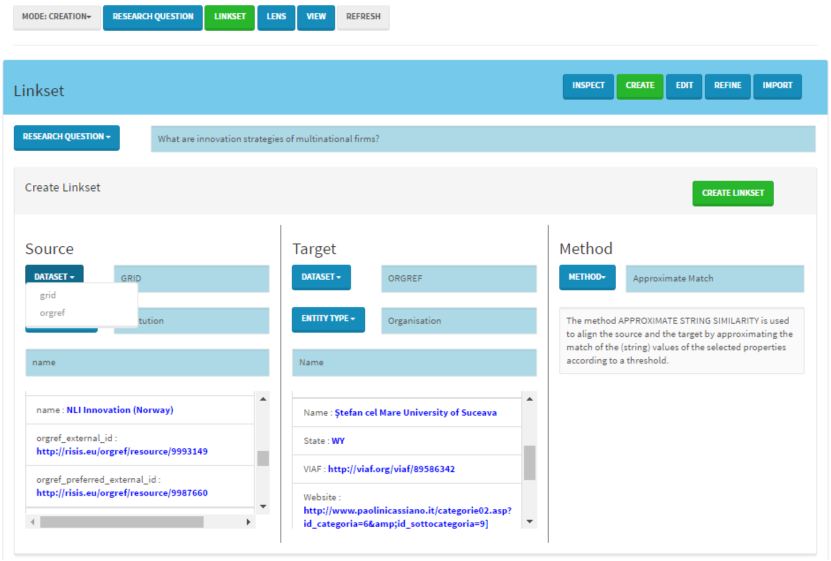
[7] https://docs.google.com/document/d/1JoJM7VF_ZaaAPbSjtgpydzRDYLvr-tROzhITGj0cH3w

[8] see an screencast of the SMS Google spreadsheet add-on at https://youtu.be/qZGDD5RN7pI

[9] see an screencast of the geo-enrciher UI at https://youtu.be/PFalWjluMR8

[10] See https://grid.ac/

[11] See http://www.orgref.org/web/download.htm



What are innovation strategies of multinational firms?

GRID

tution

name

name : NLI Innovation (Norway)

orgref_external_id : http://risis.eu/orgref/resource/9993149

orgref_preferred_external_id : http://risis.eu/orgref/resource/9987660

ORGREF

Organisation

Name

Name : Ştefan cel Mare University of Suceava

State : WY

VIAF : http://viaf.org/viaf/89586342

Website : http://www.paolinicassiano.it/categorie02.asp?id_categoria=6&amp;id_sottocategoria=9]

Approximate Match

The method APPROXIMATE STRING SIMILARITY is used to align the source and the target by approximating the match of the (string) values of the selected properties according to a threshold.

Fig. 4: An screenshot of the SMS data linking UI.

shown in Figure 4, SMS exposes an intuitive UI[12] to allow end-users create their own lenticular lenses available at `http://lenticular-lens.risis.eu`.

3 Use Cases

In order to demonstrate how the SMS platform can be used for research, we describe several use cases at `http://sms.risis.eu/usecases` . The use cases demonstrate different features of the platform in connection to addressing certain challenges covering topics such as investigating network structure of research organisations, browsing research data for temporal evolution of higher education, analyzing the geography of innovation and the structure of research portfolio and predicting Leiden Ranking from University environment factors.

References

1. M. Hildebrand, J. van Ossenbruggen, and L. Hardman. /facet: A browser for heterogeneous semantic web repositories. *ISWC*, pages 272–285, 2006.
2. A. K. Idrissou, A. Khalili, R. Hoekstra, and P. V. den Besselaar. Managing metadata for science, technology and innovation studies: The RISIS case. In A. Adamou, E. Daga, and L. Isaksen, editors, *WHiSe*, volume 1608 of *CEUR Workshop Proceedings*, pages 15–20. CEUR-WS.org, 2016.
3. A. Khalili, A. Loizou, and F. van Harmelen. Adaptive linked data-driven web components: Building flexible and reusable semantic web interfaces. In *ESWC*, volume 9678 of *Lecture Notes in Computer Science*, pages 677–692. Springer, 2016.
4. E. Oren, R. Delbru, and S. Decker. Extending faceted navigation for rdf data. In *International semantic web conference*, volume 4273, pages 559–572. Springer, 2006.

[12] see an screencast of the linking UI at `https://youtu.be/CcffBlCBF54?list=PLo4YbUaRFSnwJ9XJvp6rlIMsaw_rfKT9C`