RISIS RESEARCH INFRASTRUCTURE FOR RESEARCH AND INNOVATION POLICY STUDIES

Semantically Mapping Science (SMS) Platform: Documentation

Peter van den Besselaar[#], Ali Khalili*, Klaas Andries de Graaf*,

Al Idrissou[#], and Frank van Harmelen*

Vrije Universiteit Amsterdam, The Network Institute,

Department of Organization Sciences,

* Department of Computer Science

Table of Contents

1. Summary

In this deliverable we describe the SMS (Semantically Mapping Science) data integration platform (http://sms.risis.eu), the technical core within the RISIS data infrastructure for *Science. Technology and Innovation Studies* (STI). The aim of the platform is to produce richer data to be used in social research – through the integration of heterogeneous datasets, ranging from tabular statistical data to unstructured data found on the Web. We outline the platform's architecture and functions. There are also some example use cases mentioned to show how the platform enables data integration in practice.

2. Introduction

Up to now, STI 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, OECD STI indicators, etc., and therefore deploying only a few variables – determined by the data available. However, progress in the STI research field 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. To enable that, combining and integration of STI data and beyond is needed – in order to exploit the huge amount of data that are 'out there' in an innovative and meaningful way. That is why the core of the SMS platform is the conversion of different datasets in a standard open format: from tabular data, text data and web data to RDF (Resource Description Framework) data.

How to capture **new insights** by integrating data from multiple heterogeneous data sources in the STI domain?

This emphasis on data integration is also visible in other research fields. That enables us to build a data infrastructure partly by reusing existing tools. Within the RISIS project we develop the *SMS platform for data integration and data enrichment* by combining those existing tools with specific tools newly developed for the STI field. In this report, we first describe the architecture and then the different functions that the SMS platform offers.

3. Conceptual Model

SMS platform at its conceptual model employs an entity-centric approach to interlink heterogenous datasets in the STI domain. As shown in Fig 1, 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 SMS infrastructure. A demo on how linked entities work on SMS, is provided at

https://youtu.be/rQxgGXQccqw?list=PLSBPxopOi20XPOn1sGBthbNtXIUOqM_4b .

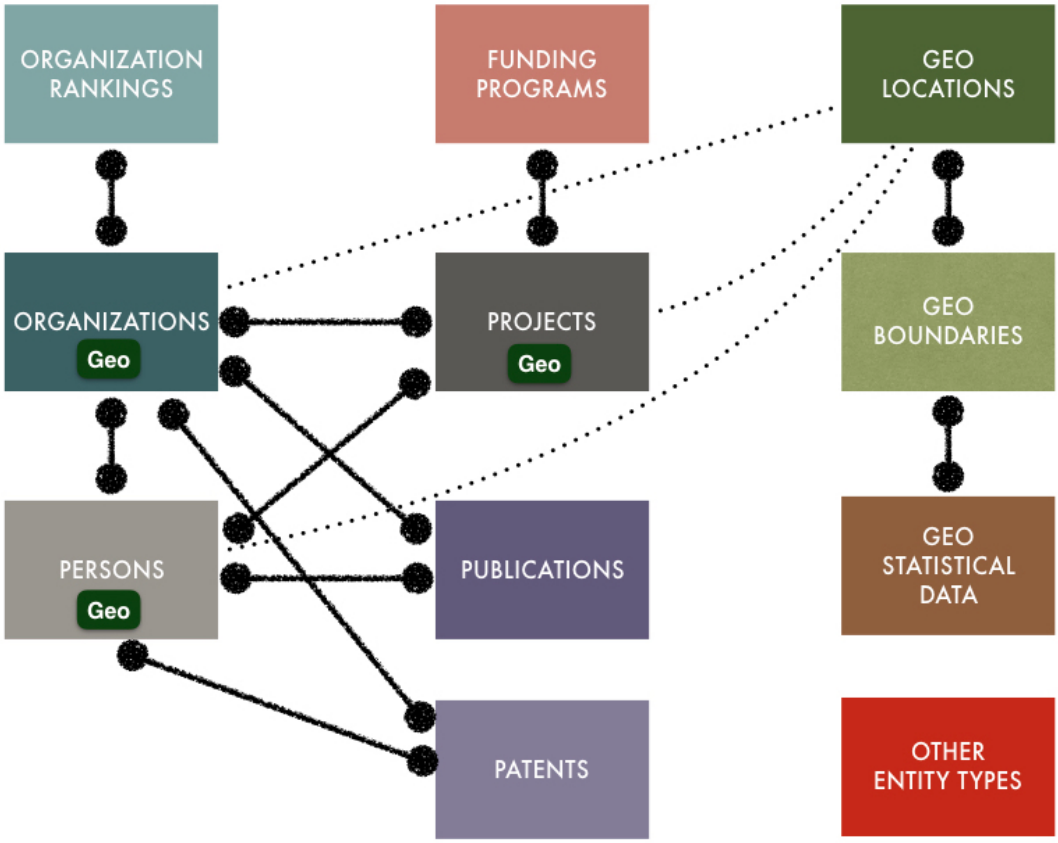


Fig 1. Main entity types supported by SMS platform

Fig 2 shows the list of currently linked datasets on SMS mapped to their corresponding entity type. There are three types of datasets in the list: *public datasets* which can be accessed by anyone, *private datasets* which are only accessible by certain users, *subscription-based datasets* which could be accessed by users who have paid subscription to data.

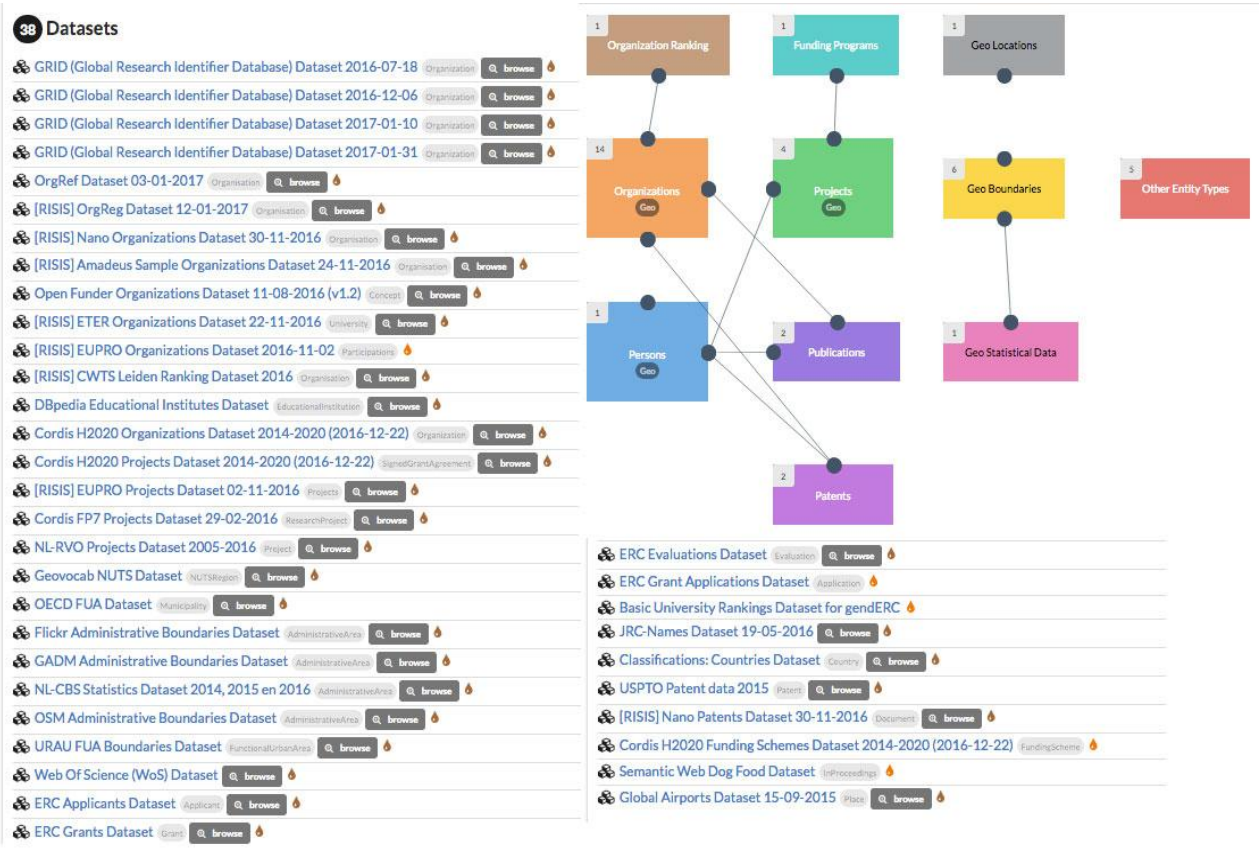


Fig 2. List of existing datasets on SMS grouped by their corresponding entity types

Public datasets are e.g., GRID, OrgRef, ETER (RISIS), OrgReg (RISIS), Leiden Ranking (RISIS), Cordis. Private datasets that require a subscription with the owner are e.g., the (links with) Patstat, Amadeus, and WoS. Another category of private datasets several of the RISIS data, such as EUPRO and Nano, which require permission of the data owner. Finally there is the possibility to link for an individual research project confidential data, as in figure 1 the ERC grant applications dataset.

4. Technical Architecture

As shown in the Fig 3, the SMS platform has a layered design; from data sources (bottom) to data services and functions for end-user (top). We describe the layers starting from the bottom layer and ending with the top layer in the sections below.

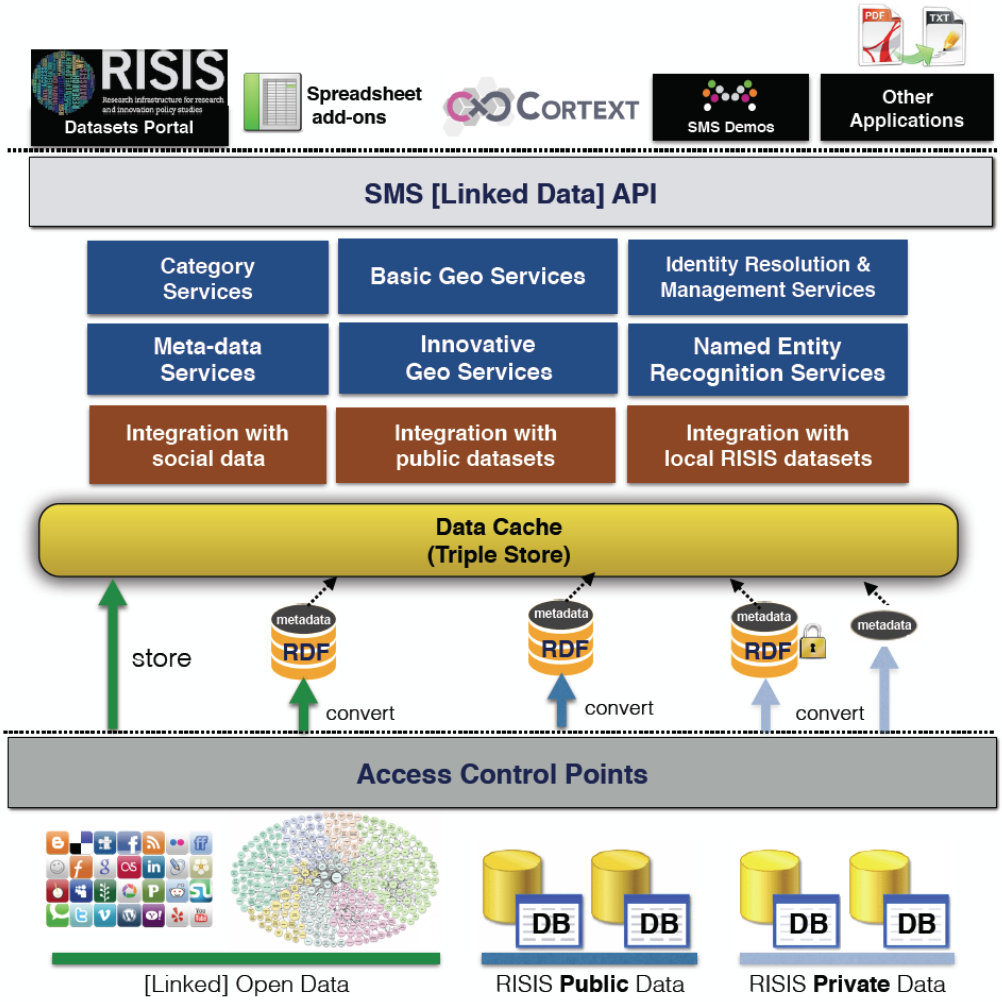


Fig 3. 3-layer Architecture of SMS Platform

We describe different components of SMS platform based on the data flow in system (as depicted in Fig 4). Data either collected from RISIS dataset repository or open data on the web, is first converted to Linked Data format. On top of the created linked data, a set of Web services are provided which allow different applications to plug and take benefit of linked datasets. SMS already provides a set of applications which combine Linked Data services to address user needs. These applications allow researchers to find answers to their research questions defined on specific use cases.

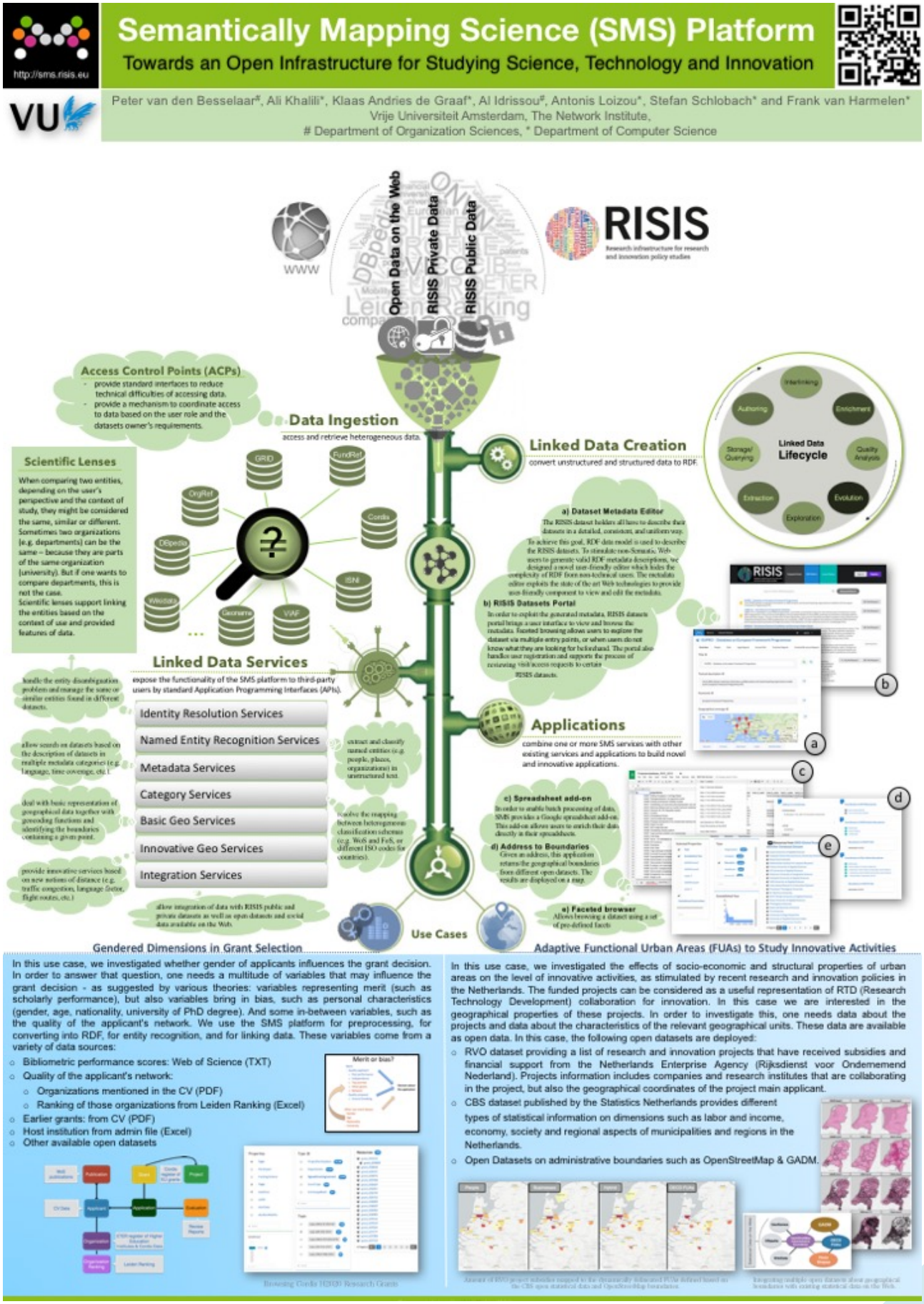


Fig 4. Data flow in the platform: from data to use

Data Ingestion

Importing data to SMS platform can be done both manually and automatically based on the "Entity types' covered by a dataset, ' Format and structure' of data and 'Data access policy' defined for data to be imported. The latter is important as not all data can be accessed by every user, and different levels of accessibility apply, depending on subscriptions and on permission of the owners of datasets.

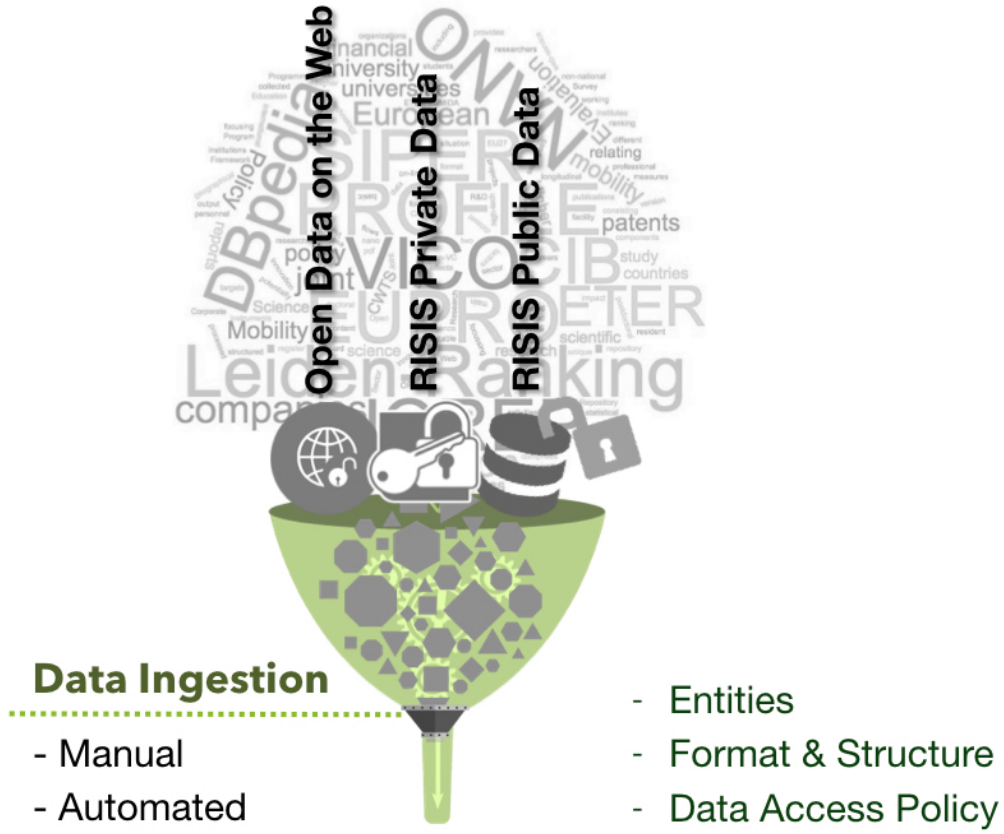


Fig 5. Data Ingestion in SMS Platform

Following questions need to be answered before importing data into SMS:

- ***What types of entities are covered by the dataset?***
  The answer to this question, helps SMS to find the potential points of linking and also to check if the conceptual model should be amended to accommodate new entity types.
- ***What is the format and structure of data to be imported?***
  The answer to this question, helps SMS to automate the ingestion process if the data format and structure are based on the standard interfaces supported by SMS.
- ***What are the data access policies?***
  The answer to this question, helps SMS to apply restriction rules when accessing the imported dataset.

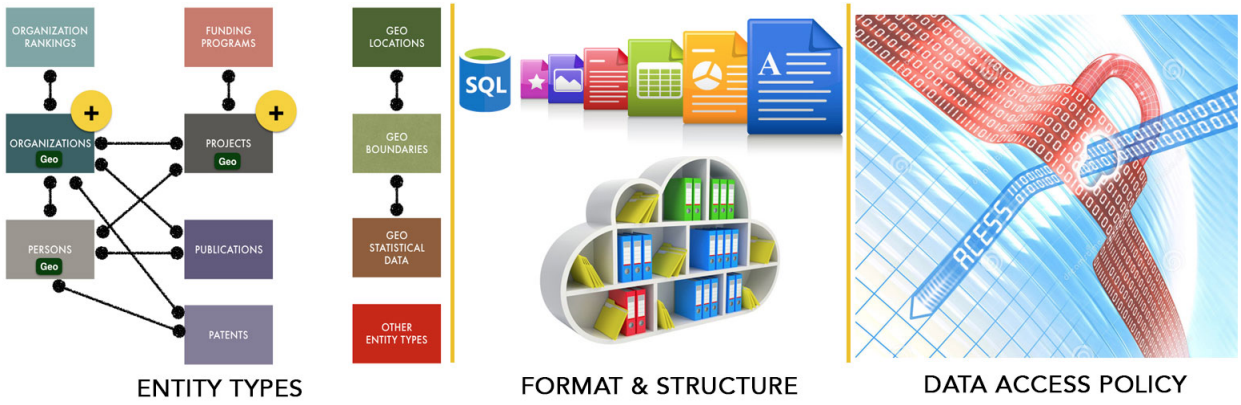
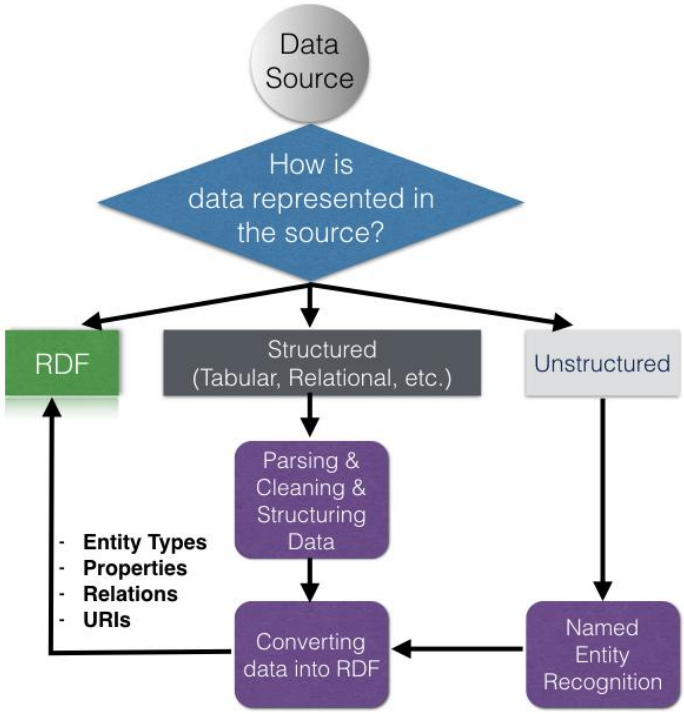


Fig 6. Important factors for data ingestion

Linked Data Creation

There are several steps followed in the lifecycle of linked data to extract and store the imported data into SMS triple store. The lifecycle starts by a basic (syntactic) conversion of data to RDF format without applying any specific vocabularies. This basic conversion is then enriched by applying several linking and enrichment services. Different services and scripts are used to convert unstructured and structured data to RDF. Techniques such as Named Entity Recognition (discussed later in the document) can be employed to extract named entities from textual content. A concrete example is recognizing research institutions and universities in a researcher's CV (Curriculum vitae), using named entity recognition by linking the CV to databases with background knowledge such as DBpedia.



For structured content, the tool will be selected based on the format. For example, OpenRefine[1] can be used to convert spreadsheet data to RDF.

[1] http://openrefine.org/

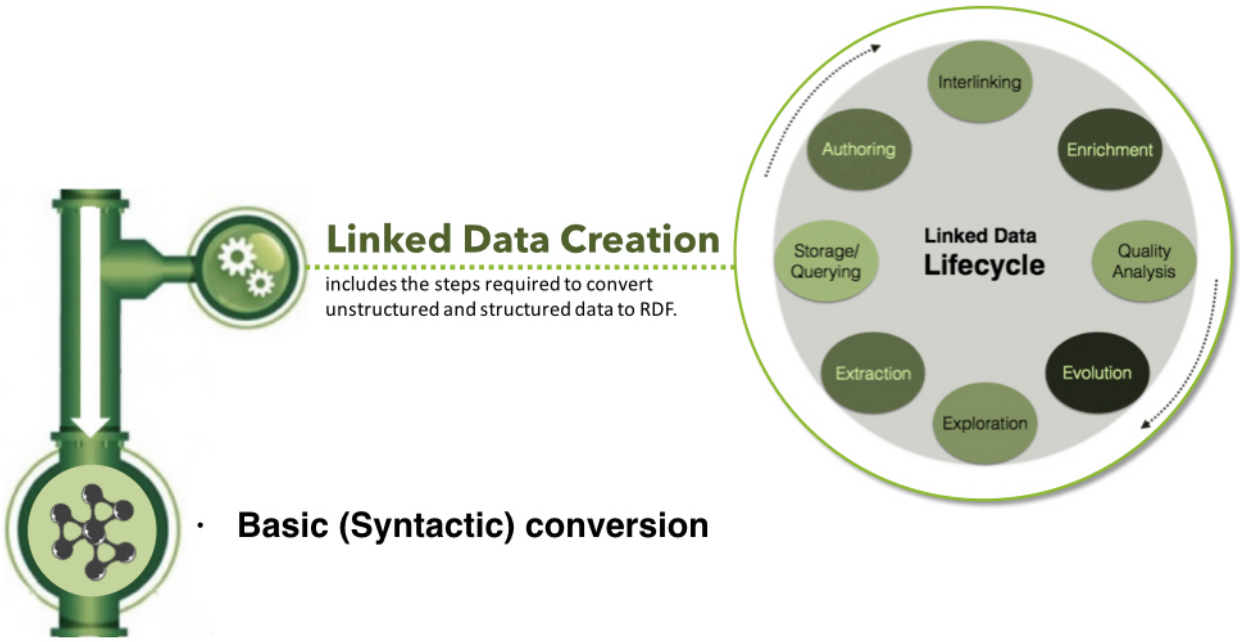
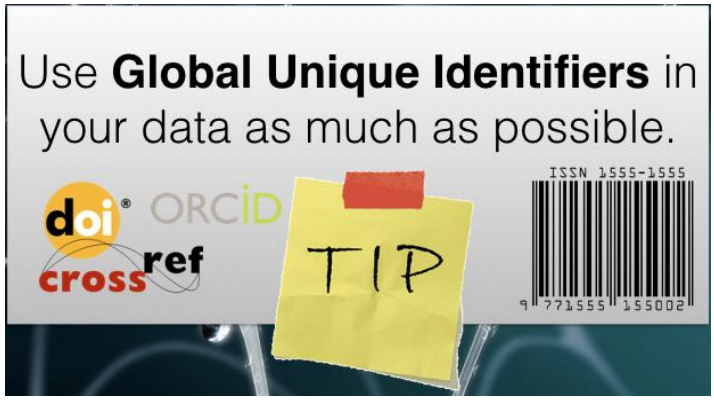


Fig 7. Linked Data creation

Data Linking and Scientific Lenses



Data linking is the process of creating a relationship between entities that meet preset conditions. If global unique identifiers for entities are available, the linking becomes straightforward. If not, a variety of techniques can be used, from (fuzzy) string matching to deploying attributes available in the different databases. In the link data service that we provide, we emphasis on providing contextual information that help eliminating ambiguity after a relationship between entities is established, and enables re-use. For instance, the GRID[2], OrgRef[3], and EUPRO[4] datasets describe organisation entities across various countries, including both public and private research organisations. All of these datasets refer to the "Minnesota Mining and Manufacturing Company" (3M), a large multinational organisation with a substantial patent portfolio. The GRID dataset distinguishes between 3M(United States) and 3M(Canada), while the OrgRef dataset only refers to the single entity 3M. To study these organizations, they need to be aligned across these datasets whenever they are the same. But what does "the same" mean? Suppose one study aims to compare organizations at a global level, whereas a second compares organizations across countries. In the first setting, all occurrences of '3M' in the datasets are considered the same. In the second study, the Canadian and U.S. branches of '3M' are to be considered separately.

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

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

[4] See http://datasets.risis.eu/

In our approach to data linking, we first provide a network of interlinked entities through linksets. These linksets are generated using basic similarity metrics such as exact string similarity, approximate string similarity and geo-similarity. The goal of these linksets is to serve as “lego pieces”, easy for users to combine or modify them to their liking to answer a particular research question. Combining or modifying linksets is made possible using operations such as UNION, TRANSITIVITY or INTERSECTION. The result of a manipulation over one or more linksets is a lens, which stands as a user view over the data.

We propose to enable users to make an informed choice over alignments produced by existing tools. This modifies the generic problem into choose and modify. Our proposal is to reuse existing tools for generating correspondences of as the basis of interlinking.

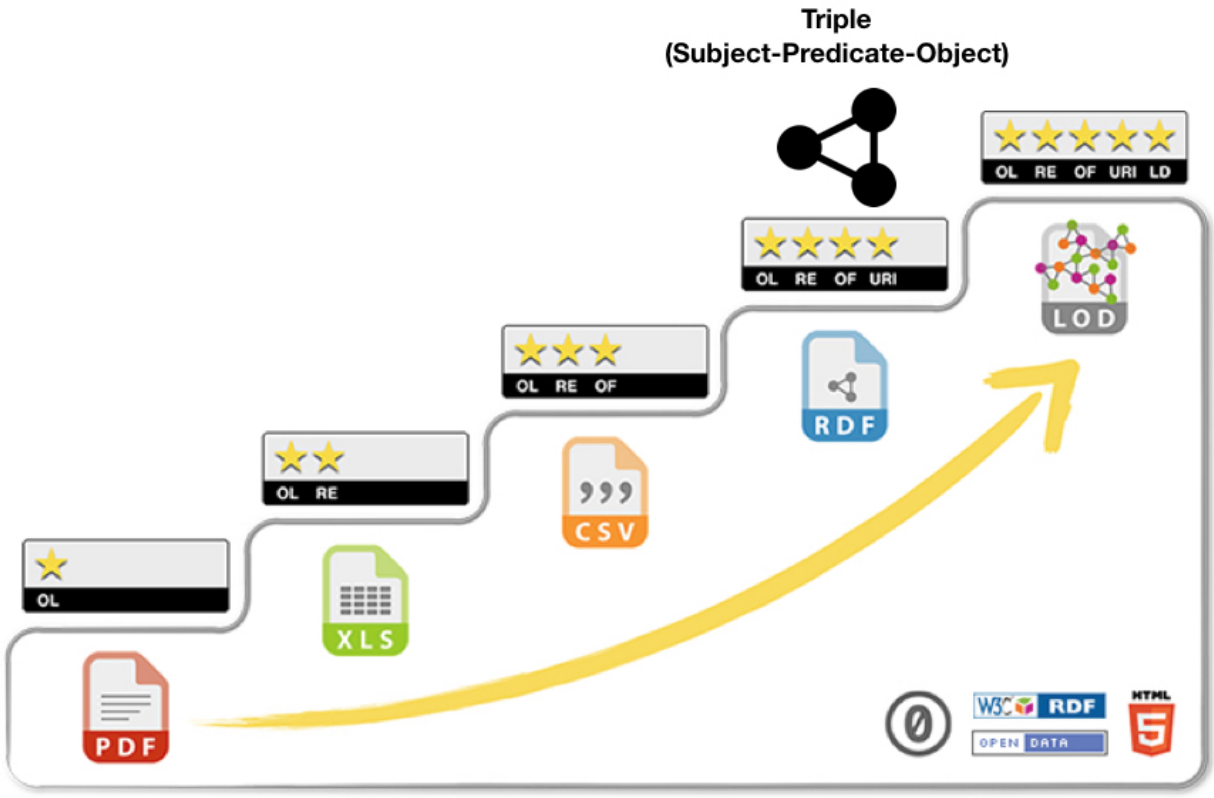


Fig 8. From raw data to Linked Data with higher expressivity

Linked Data Services & Applications

SMS platform exposes a set of predefined SPARQL[5] query templates as RESTful[6] Web APIs in order to facilitate usage of the interlinked data by developers who are not familiar with the SPARQL query language. The Web services also allow better management of data access (in case authentication and authorization are needed) while monitoring the data usage for optimizing the queries and provide load balancing on the services infrastructure (e.g. due to reasons of data size and performance of the respective geospatial queries, scalability of Linked Geo Data platforms is a critical issue and needs to be dealt by distributing the services into a set of composable micro services).

An important benefit of exposing data as service is the ability to build applications which combine one or more services with other existing services and applications to build novel and innovative STI applications.

SMS uses Swagger[7] to document the APIs of the exposed Linked Data services. The full documentation of services is available at http://api.sms.risis.eu .

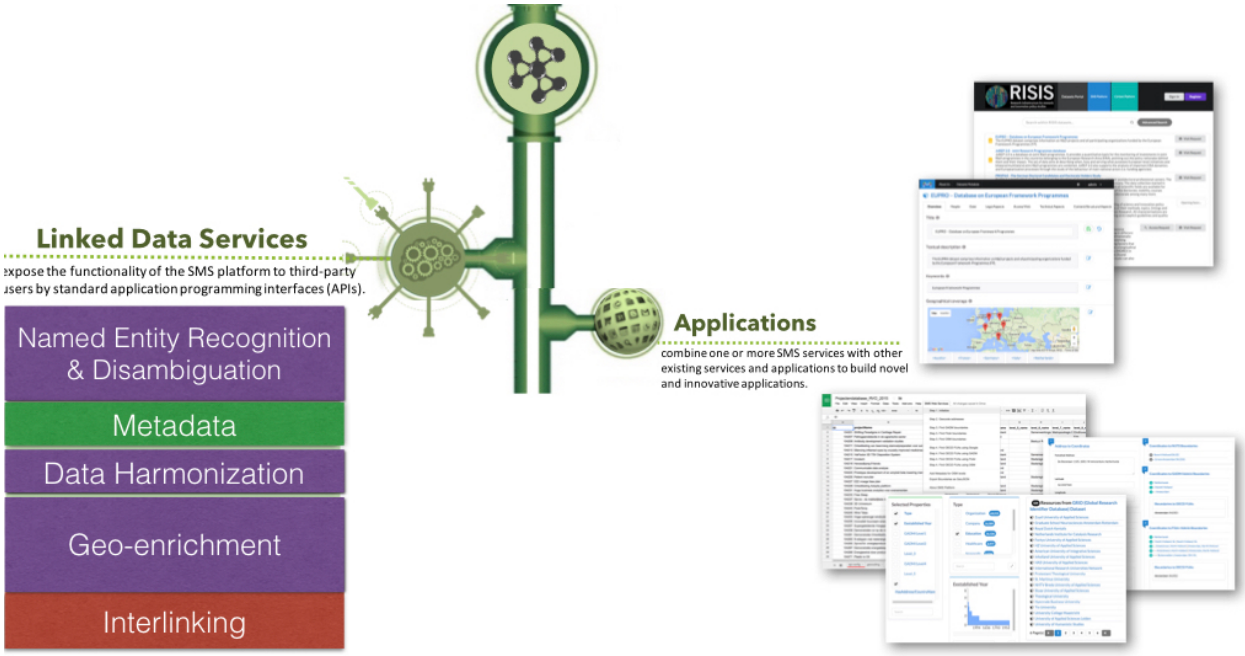


Fig 9. Category of Linked Data services with some example applications

The APIs are generally categorized into the following categories:

- Metadata Services and Applications
- Data Enrichment Services and Applications
  - Named Entity Recognition
  - Data Harmonization
  - Geo-enrichment
- Data Linking Services and Applications

[5] https://www.w3.org/TR/sparql11-query

[6] https://en.wikipedia.org/wiki/Representational_state_transfer

[7] http://swagger.io

Metadata Services and Applications

Metadata helps potential users of a dataset to decide whether the dataset is appropriate for their purposes or not. RISIS project aims to provide a distributed infrastructure for research and innovation dynamics and policies. This infrastructure 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 RISIS 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 - modelled 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. Metadata service allows dataset holders to describe their datasets in a detailed, consistent and uniform way, store the description and if needed modify the stored metadata. The metadata can also be utilized to facilitate data integration as shown below:

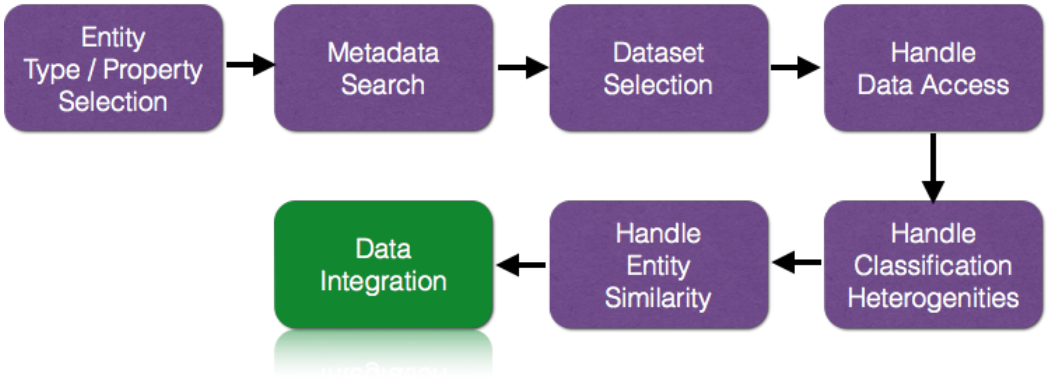


Fig 10. Importance of metadata for data integration

In order to enable end-users to easily view and edit metadata, SMS provides a metadata editor application (as shown in Figure 10) built on top of the metadata services. This application allows dataset owners to edit the metadata related to their datasets in different categories. Furthermore, for researchers interested in RISIS datasets, it provides interfaces on http://datasets.risis.eu portal to view metadata and then request to get access to the data. The following online video demonstrates how the metadata editor works:

https://youtu.be/p_2D3ydcx1U?list=PLSBPxopOi20XPOn1sGBthbNtXIUOqM_4b

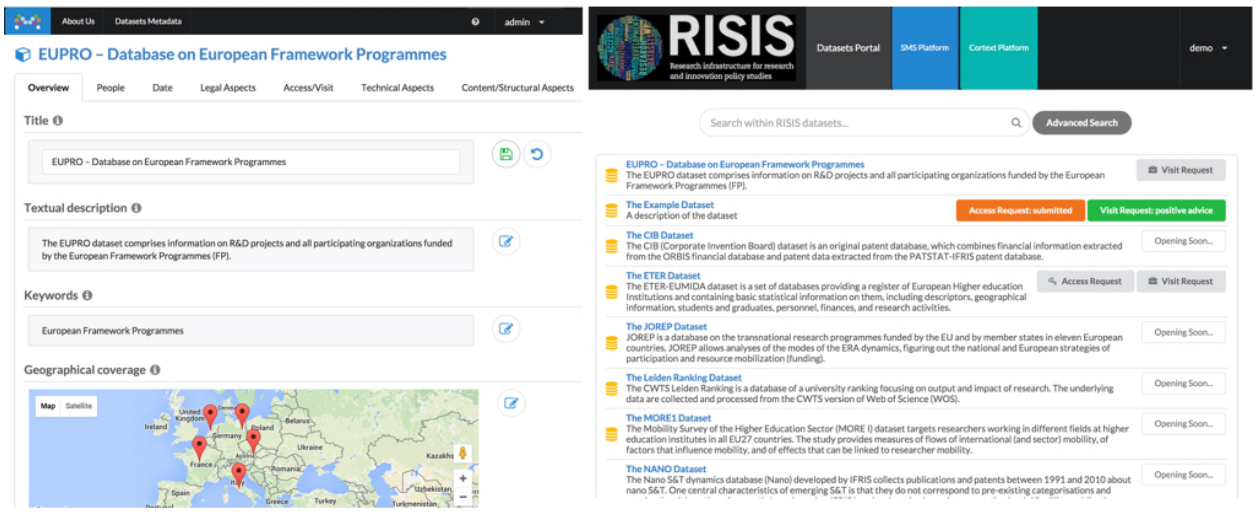


Fig 11. SMS Metadata editor used to provide description of RISIS datasets on datasets.risis.eu

Data Enrichment Services and Applications

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, entity chunking 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.[8] 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).

[8] https://en.wikipedia.org/wiki/Named-entity_recognition

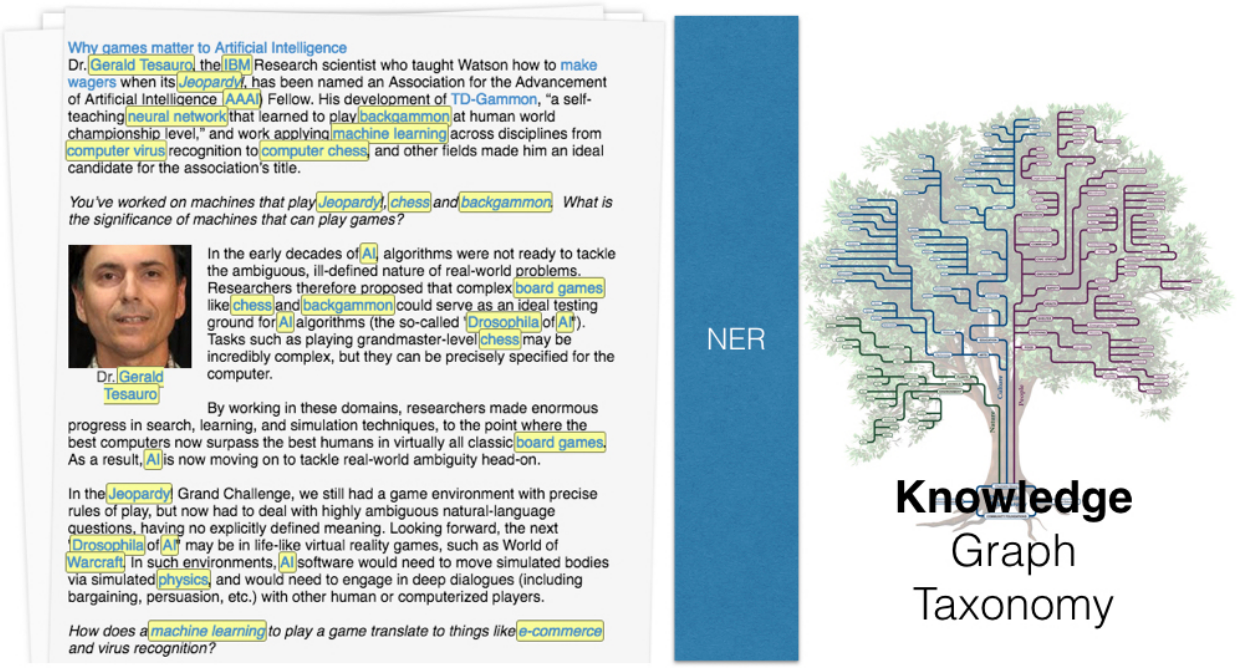


Fig 12. Named Entity Recognition

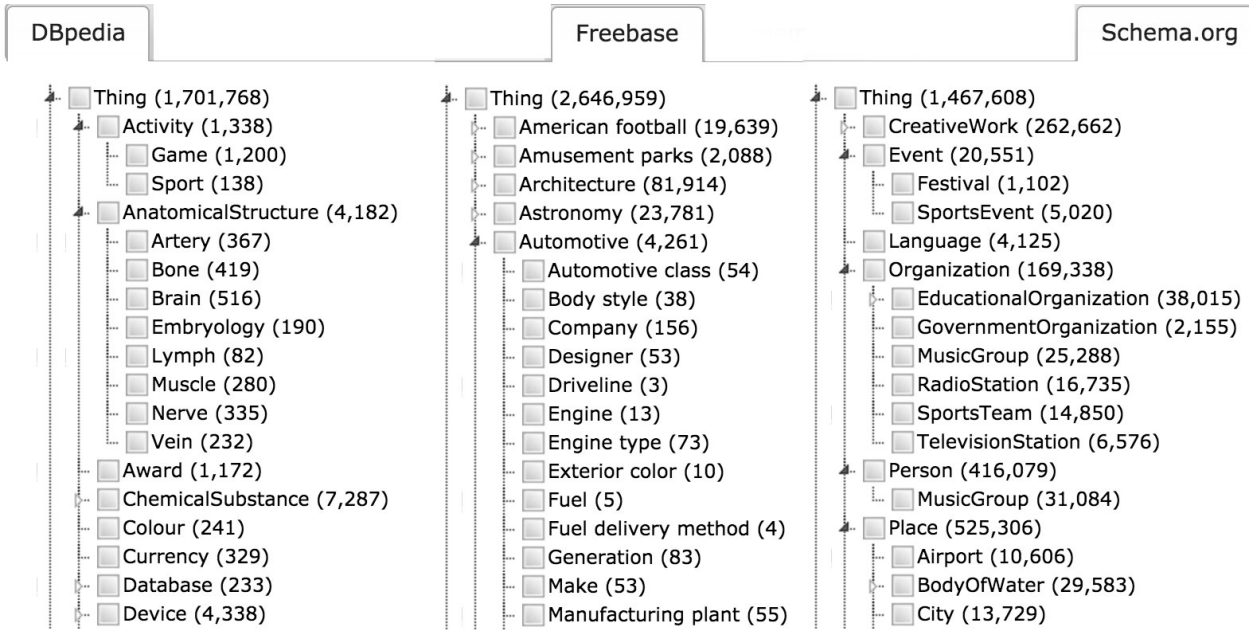


Fig 13. Different knowledge graphs used for NER

By default, SMS employs DBpedia Spotlight[9] service for NER. However, any arbitrary NER service can be plugged into SMS NER service as long as the output of service is reconciled to SMS named entities

[9] https://github.com/dbpedia-spotlight/dbpedia-spotlight

annotation model. DBpedia Spotlight automatically annotates mentions of DBpedia (structured information extracted from Wikipedia) resources in text. The extracted entities map to a taxonomy of general knowledge (as shown in Fig 13, DBpedia, Freebase and Schema.org ontologies) which helps users to better browse and analyze a dataset taking a particular domain of interest.

SMS faceted browser allows users to browse an annotated dataset by combining the background knowledge extracted from named entities with the inherent attributes of a dataset. The following online video demonstrates how the faceted browsing of NEs works:

https://youtu.be/H76afW67qy8?list=PLSBPxopOi20XPOn1sGBthbNtXIUOqM_4b

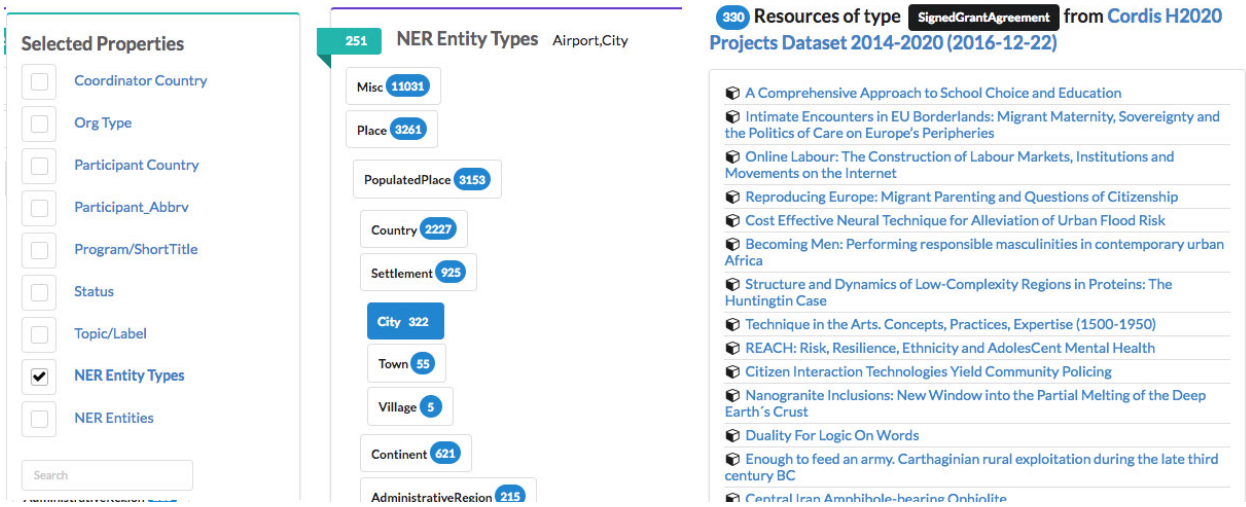


Fig 14. Faceted browsing of data using extracted named entities

Data Harmonization

The goal of SMS data harmonization service is to improve the quality and expressiveness of a dataset by enriching it with an existing standard classification. The harmonized datasets can be more easily interlinked with other datasets. For example, with regards to geo data, data can be enriched by adding HASC (Hierarchical Administrative Subdivision Codes) or ISO 3166 country codes. Or with regards to publication/patent data, using FoS (Field of Science), WoS (Web of Science) or IPC (International Patent Classification) classifications.

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 system

is based on a series of open geo-resources, such as GADM[10], OpenStreetMap[11] and Flickr[12] geotagged data. By integrating these geo-resources, the service can give for an entity's address the geo-location up to 11 different levels. We illustrate this with an example of a service to determine the geographical location if one knows an address (or even only an organization name). As shown in Figure 15, in the top right part of the screen the address for "Vrije Universiteit Amsterdam" is inserted, and the application has as output various maps and, in the bottom right, the geo-characterization of the inserted address at eleven levels.

Figure 15 shows the various administrative boundaries for the geocoded address. A simple address-to-boundary application is available, which can be used to check different geo-boundaries used with their corresponding metadata:

at http://sms.risis.eu/demos/geo/addressToAdmin

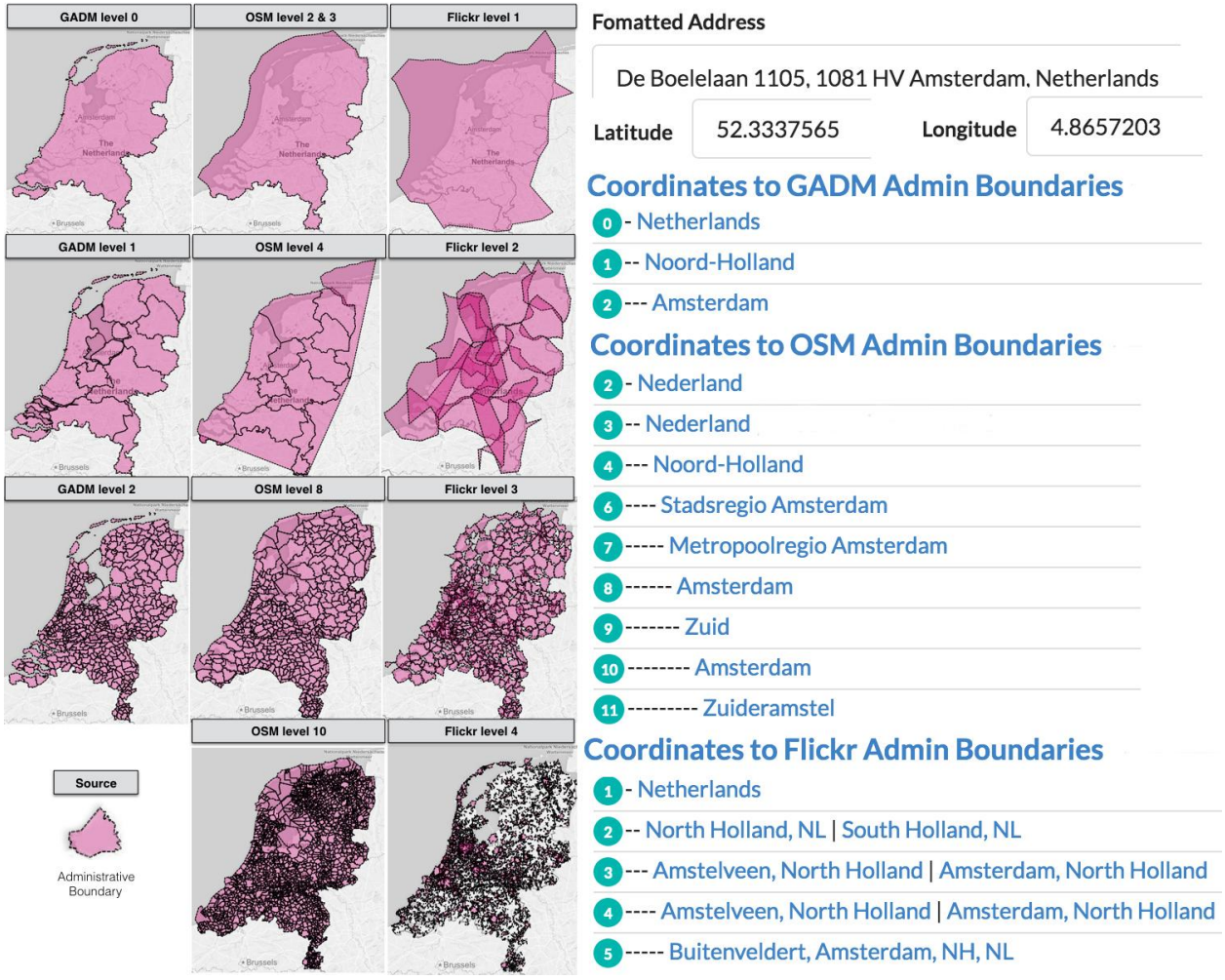


Fig 15. Geo boundaries extracted from open resources on the Web

[10] Database of Global Administrative Areas: http://www.gadm.org

[11] http://www.openstreetmap.org

[12] http://www.flickr.com/services/shapefiles/2.0/

With regards to geo-enrichment, SMS provides the following main categories of services:

- ➔ Geocode a given address.
- ➔ Find administrative boundaries containing a given point.
- ➔ Find metadata and details of a given administrative boundary.
- ➔ Find (multi-)polygon shapes of a given administrative boundary.
- ➔ Find Functional Urban Areas (FUAs) related to a given administrative boundary.
- ➔ Connect administrative boundaries to selected statistical data.

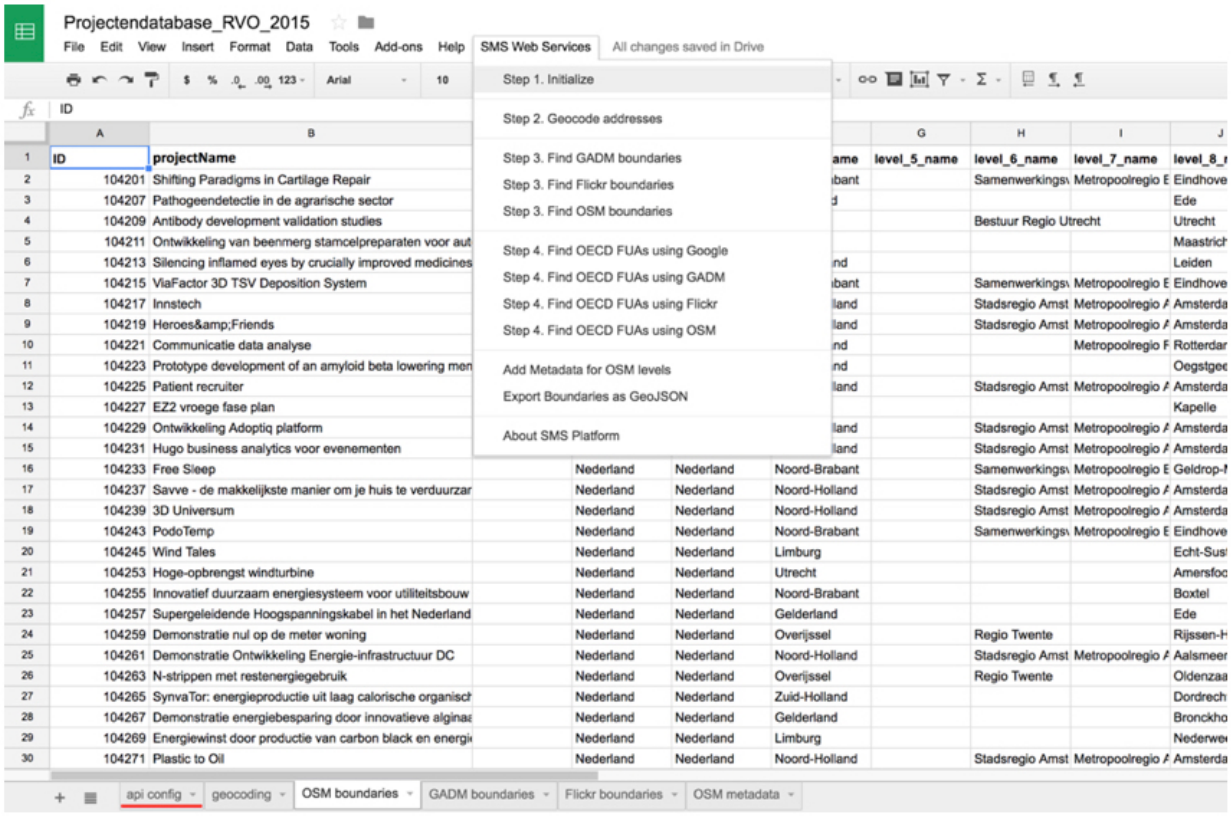


Fig 16. Google Spreadsheet add-on

One practical application we built for batch processing of addresses is a Google spreadsheet add-on (see Figure 16) which chains Google Geocoding API with our PointToAdmin and AdminToFUA 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[13]. The following online video tutorial demonstrates how to use our Google spreadsheet add-on:

https://youtu.be/qZGDD5RN7pI?list=PLSBPxopOi20XPOn1sGBthbNtXIUOqM_4b

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. For a dataset that does not include geo

[13] https://carto.com/

coordinates, addresses will first get automatically geocoded by Google Geocoding API to include longitudes and latitudes. For datasets that are already geocoded, the SMS boundary services will be immediately applied to extract the container boundaries in different levels for existing open geo boundary sources.

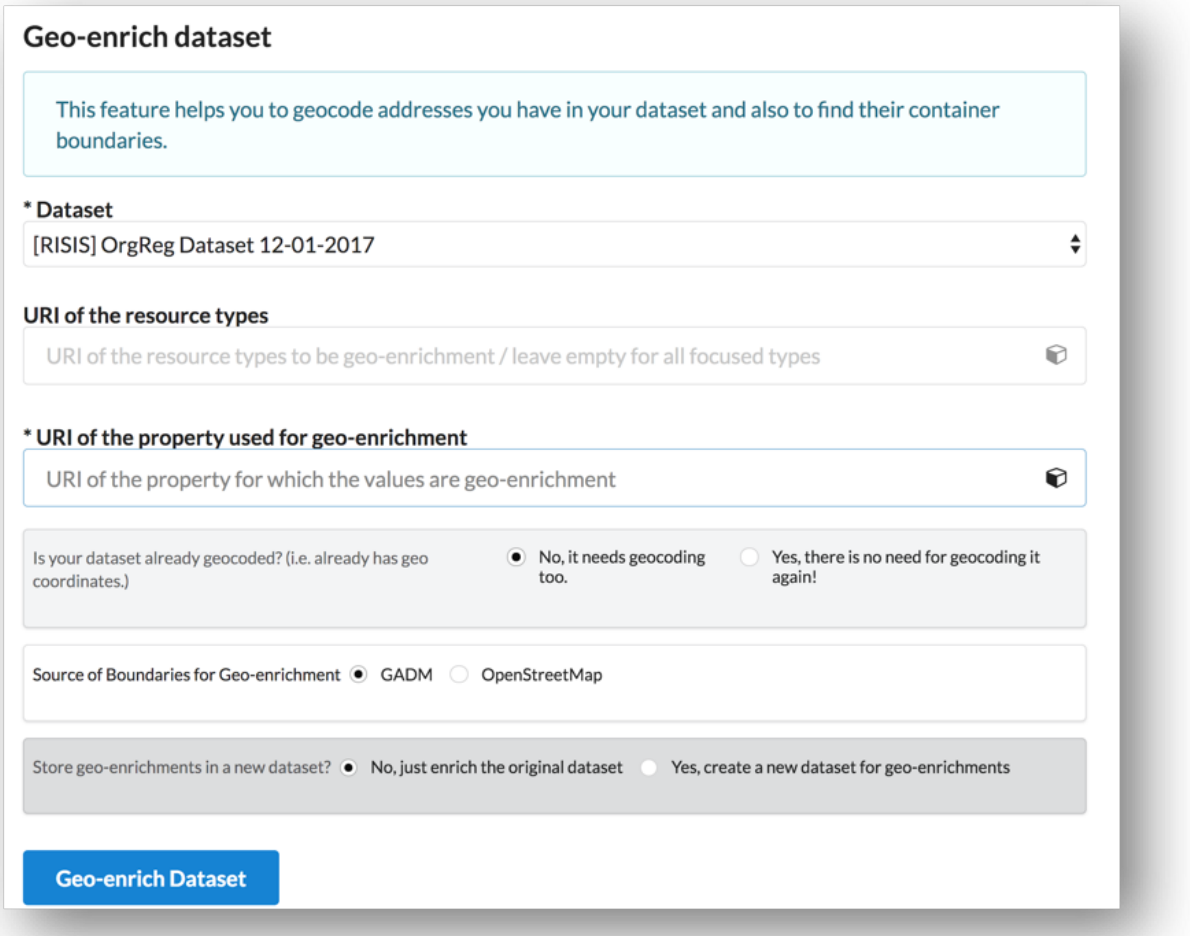


Fig 17. Linked Data geo-enrichment UI: configuration

The result of geo-enrichment can be stored either directly in the original dataset or in a separate dataset with links to original dataset. The interactive user interface allows users to see in real-time the geo-enriched entities on a map with their extracted geo boundaries.

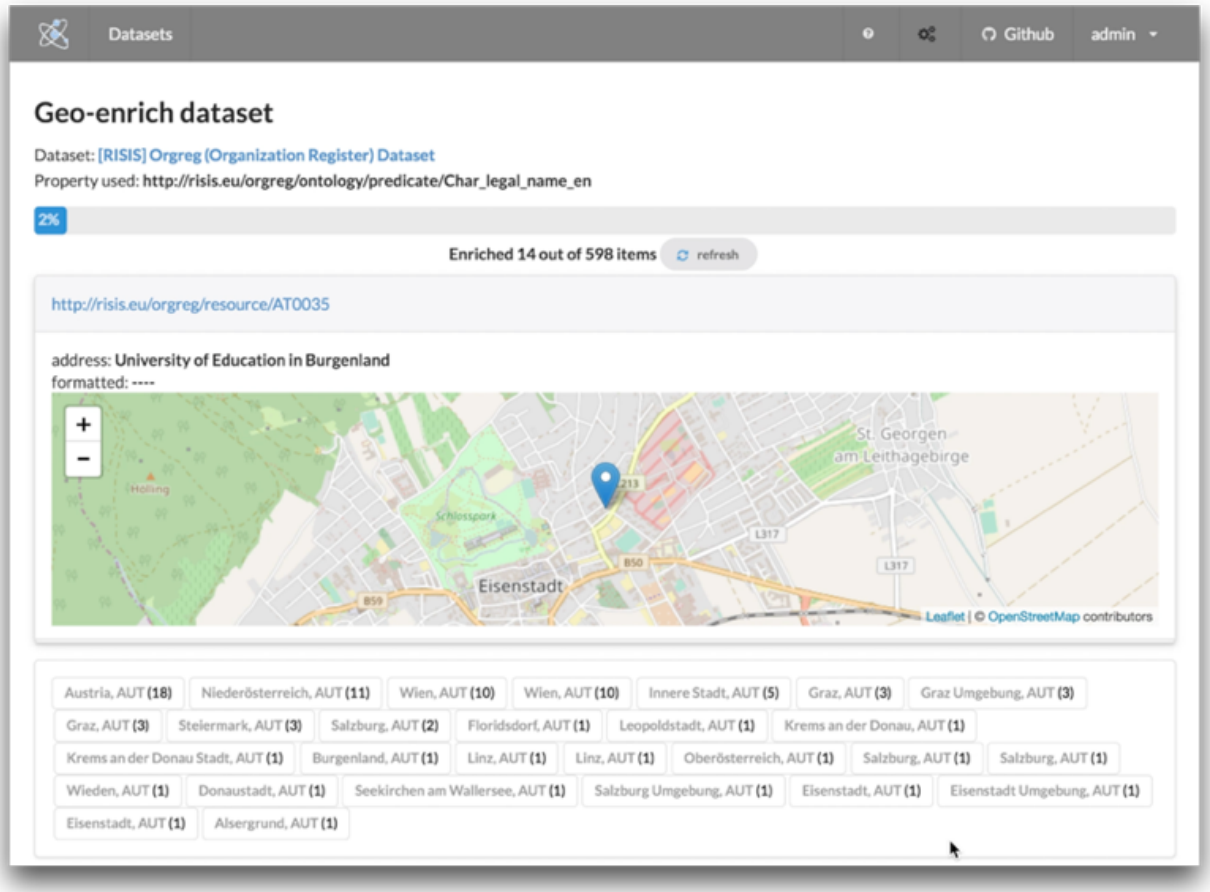


Fig 18. Linked Data geo-enrichment UI: progressive annotation

For the geo-enriched datasets, users can use the SMS faceted browser to display the entities within the datastes on an interactive map and combine geo-data with other structural attribute of the datasets to facilitate browsing the datasets.

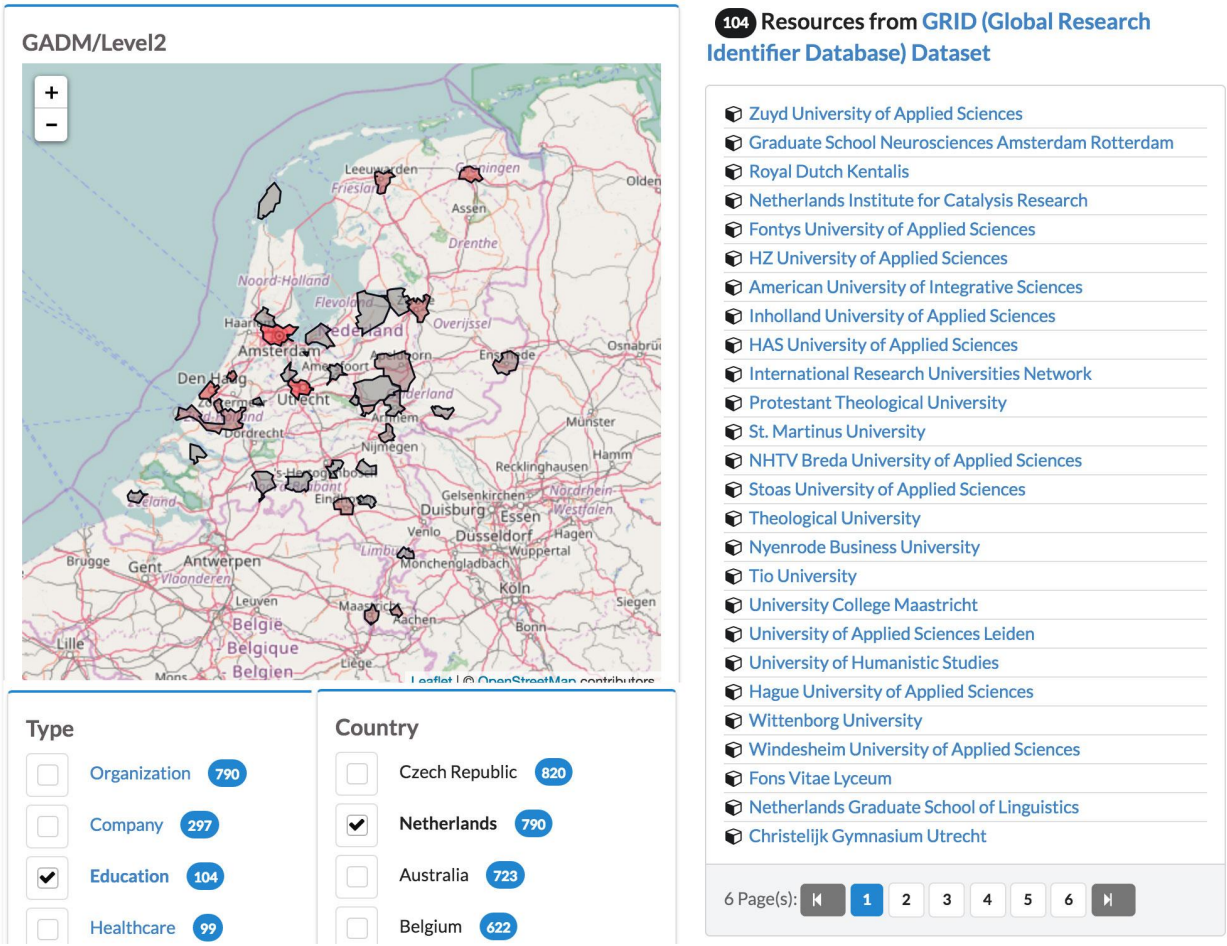
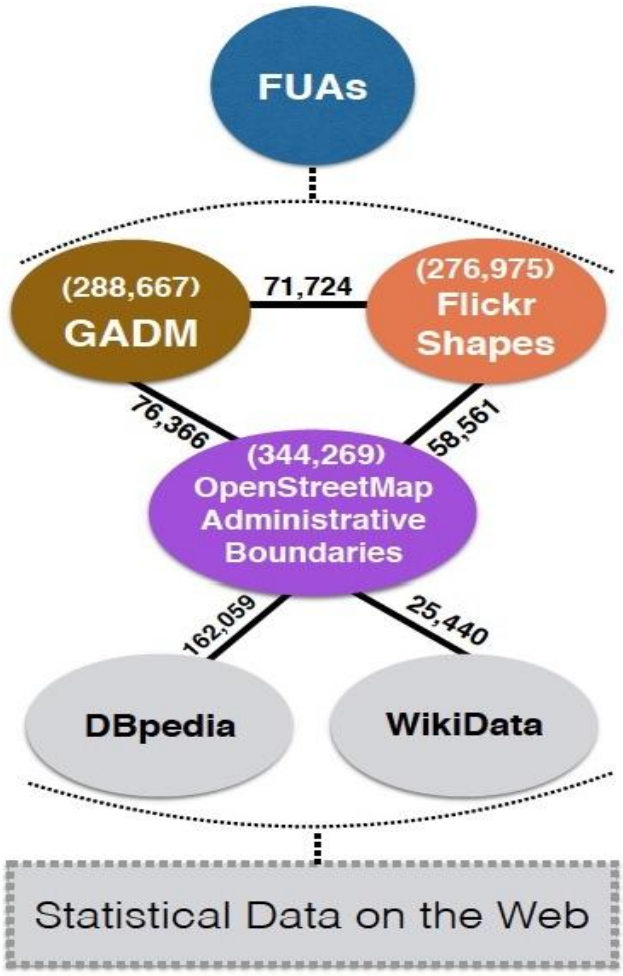


Fig 19. Faceted browsing of geo-enriched data

The following online video demonstrates how the Linked Data geo-enrichment service works :

https://youtu.be/FFy4-Zlt_ak?list=PLSBPxopOi20XPOn1sGBthbNtXIUOqM_4b

As another application, SMS proposes a Linked Data approach and implementation which combines openly available spatial and non-spatial resources on the Web to more flexibly classify urban areas. We have already interlinked several datasets related to open geo-boundaries. Users can choose an existing statistical dataset which provides data on certain levels of administrative boundaries and combine it with SMS linked geo data to create a new notion for urban areas. In the section related to use cases, we bring one example of delineating an adaptive urban area.



Data Linking Services and Applications

Before a user can obtain a view over the data of interest, he is to interact with our services. All his/her interactions are of value to the other users in the sense that those actions are documented for others to reuse, modify for different purposes. User interactions include:

- ❏ Mapping between research question, entity-types and datasets
  This enquires about how the research question relates to entity-types hence, datasets that describe those types of interest.

- ❏ Alignments used to generate linksets.
  Here, an explicit description of how to align datasets is required from the user.

- ❏ Lens or user view over the data
  The user provides a complete description of how she likes the data to be integrated.

- ❏ The design of a view
  The user submit the set of properties that are of interest to answer her research question.

- ❏ Link validation
  The service requires the user to confirm or reject each correspondence created between entities. The justification of the rejection or validation of a link is asked from the user. The later data is intended to help other users decide on their own whether or not to add a contradictory explanation of why the a previously judged "wrong" link should be reinserted for their particular task.

Fig 20 shows the steps a user has to take in order to describe and extract a view over the data of interest. For the sake of example, let us assume that the system already contains a set of linksets and lenses. For a user to start a linking activity, she needs a research question.

Based on the research question, she is requested to select the entities types of interest, the datasets that describes the selected entity types. From here on, all she needs to do is *"Select the lens for the view*" and "*Design the view*". Once the view is designed, the user uses the linking service to "*generate the view table*" that she will use for her analysis. After analysing the data, the user is to feed the linking service "*Associate the result of the analyses*" with her results (link to publication, report, website...) to finally end the started activity.

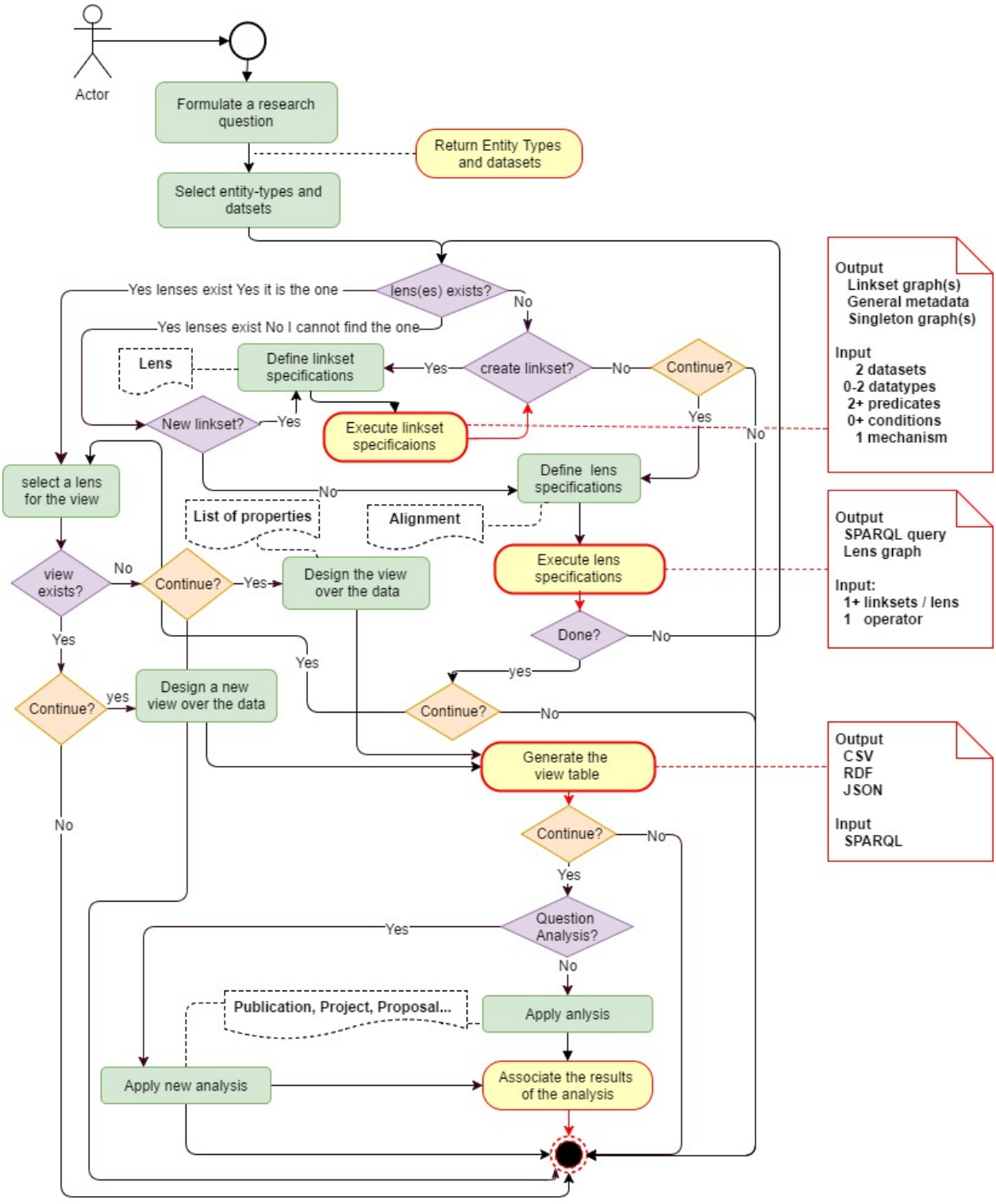


Fig 20. SMS Linking workflow

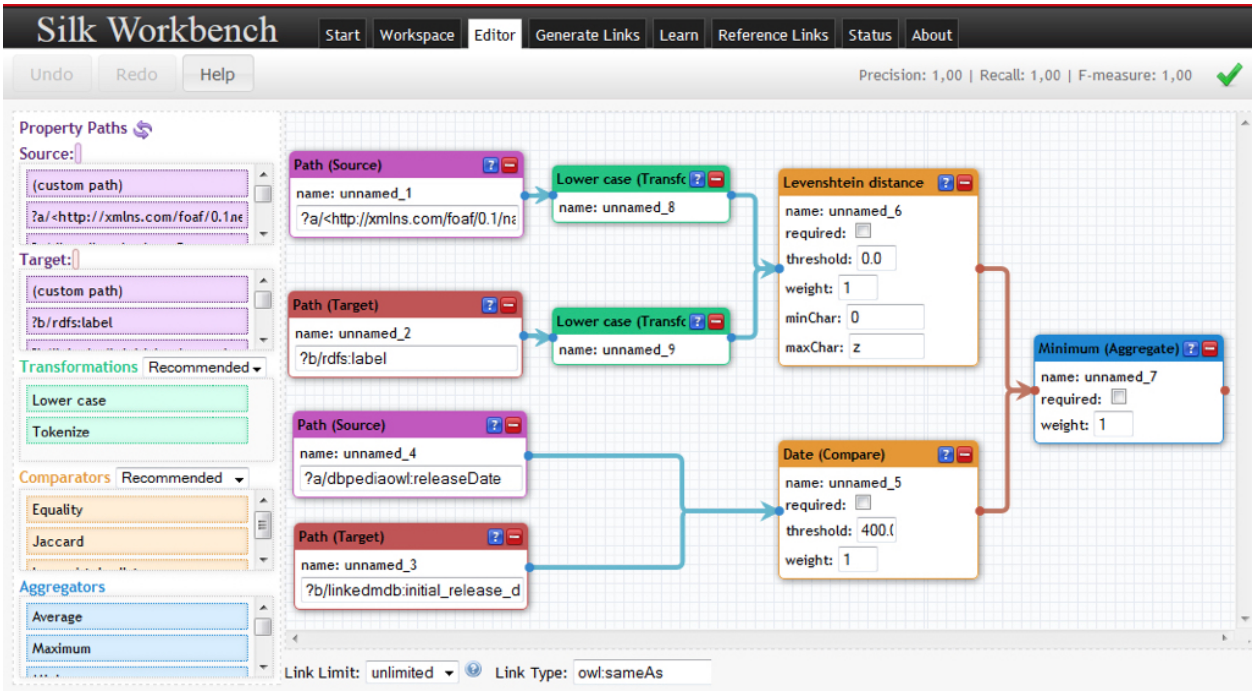


Fig 21. Data/Schema Alignment user interface

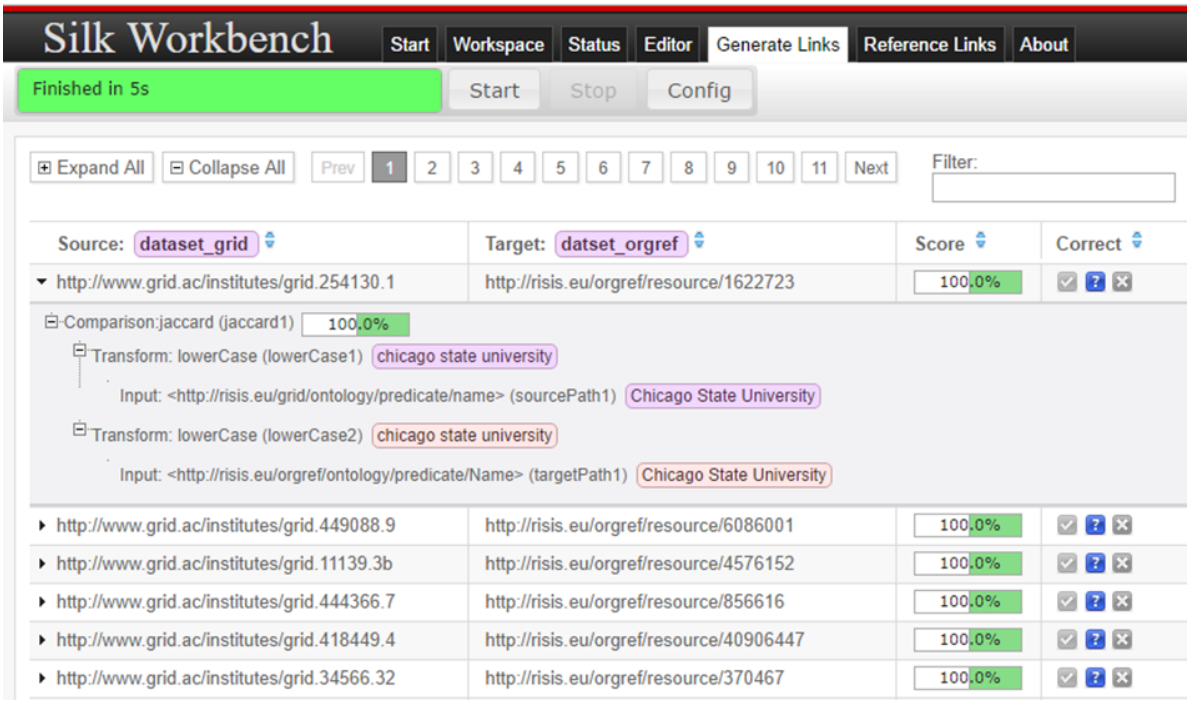


Fig 22. Results of alignments

In the linking process, other tools such as SILK[14], AGDISTIS[15], Openrefine[16] and more can be used. The figure 21. below gives an example of an alignment done with SILK prior to generating a linkset. Figure 22. shows the result of an alignment where the user can be informed about the existence of a particular link. Figure 23 shows ……

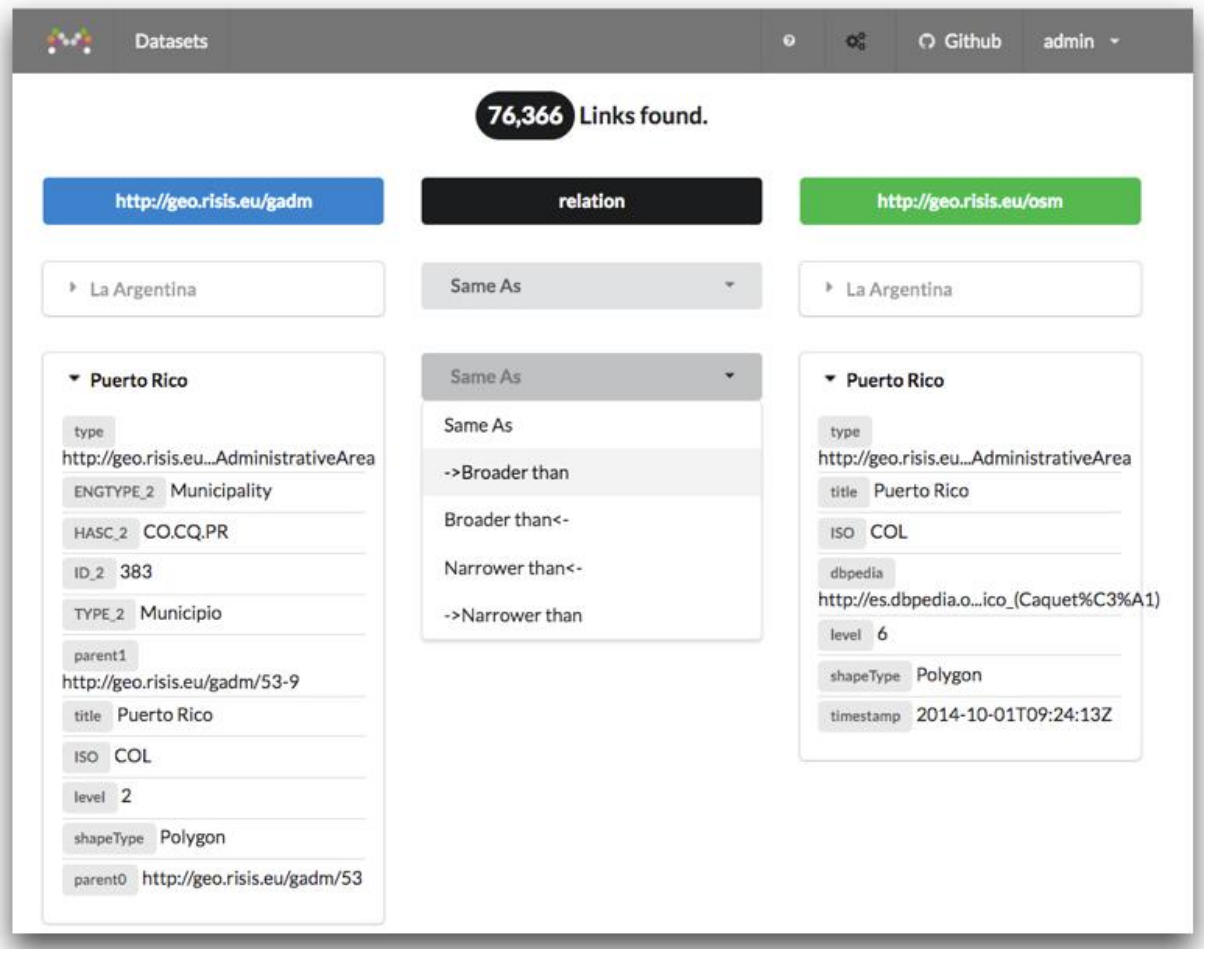


Fig 23. Linked Data validator

[14] The link data integration framework (http://silkframework.org/)

[15] Agnostic disambiguation of named entity using linked open datahttp://agdistis.aksw.org/demo/

[16] A free, open data source, powerful tool to work with messy data (http://openrefine.org/)

5. Use Cases

The use cases we describe below are stylized examples of research in order to demonstrate how the SMS platform can be used for research. So they should not be read as research reports per se.

We also do not go into an important issue about the quality and completeness of the data themselves, an important issue that will be addressed later in the project.

The examples are organized in increasing complexity. The first depends on browsing the faceted browse only, the second additionally requires the formulation of queries, for which many researchers may help. Even more help may be required in example three, as there dedicated data-linking is required. Finally the last example depends on more complex linking, and on several queries. Interested researchers may visit the SMS platform to do the more complex data processing and analysis work. See the website (www.risis.eu or www.sms.risis.eu) for information about the possibilities and support to visits.

Example 1: Using the faceted browser for analyzing change in the research/HE system.

The datastore contains many datasets with information about organizations. Assume that one is interested in structural change in higher education systems, one may want to browse through those datasets. The faceted browser can be of great help, as it enables to explore the available information in graphical form. While browsing the datasets, we find a property 'foundation year'. Selecting that property for a country, one gets the frequency of new foundations of Higher Education institutions per year (figure 24), and one sees immediately a high concentration in a two consecutive years: in 1986 and 1987 some 21 new HE institutions were founded in the Netherlands, on a total (now) of 114:. So some substantial changes in the HE system seem to have taken place.

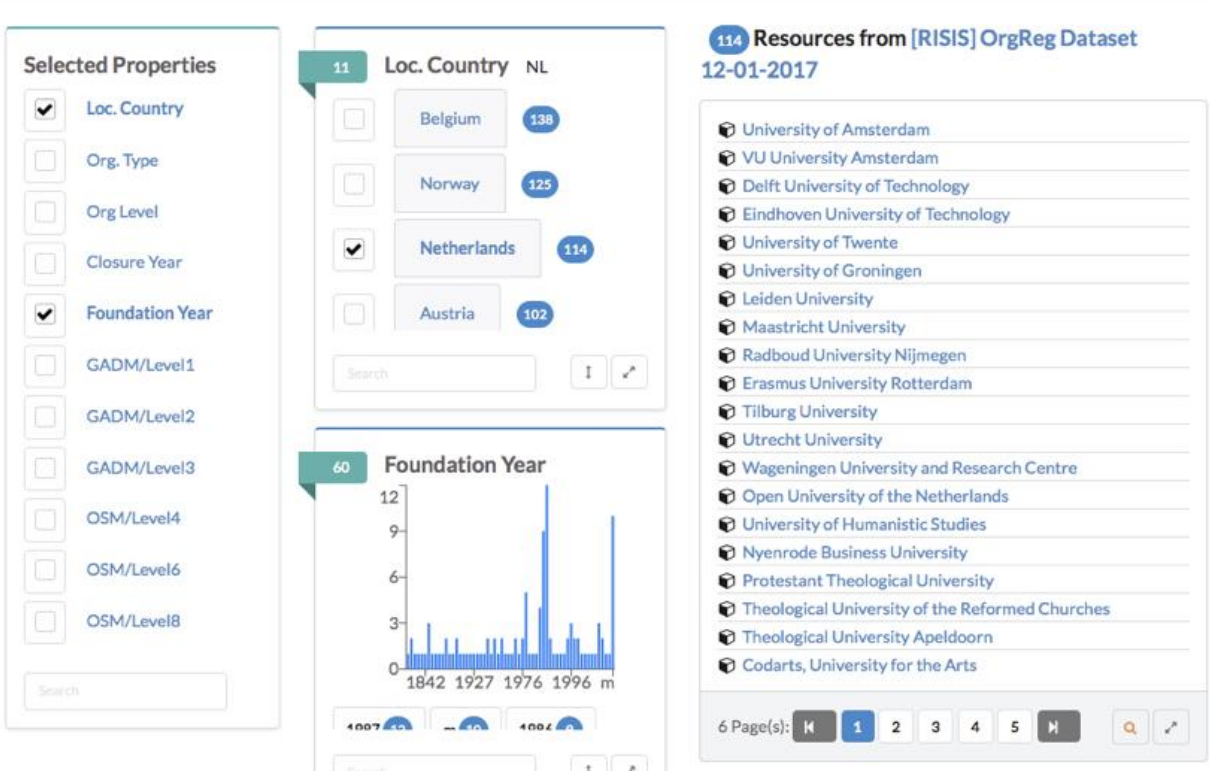


Fig 24. Foundation years HE institutions Netherlands

By selecting these two years, the list of organizations at the right side of figure 25), the screen (the 'resources') shows the names of the institutions that were founded in these two years. We can inspect the list, but also select a single institution and inspect the available information in the data store, but also more broadly on the web,as all the organizations are also linked to their website and their wikipedia page. So we do not only have much numerical data in the data network, such as numbers of students and staff, and of output, but also qualitative (textual) data for further inspection.

Looking at the various newly founded schools shows that this are all Universities of Applied Sciences, so the 'second layer' Dutch HE institutions, and one may find information on the foundation on Websites, or find contact addresses to search for further information. If one would purdue this data collection, one would find out that the new founded institutions in fact are mergers of smaller schools into very large new institutions. This indeed can be considered as a major reform of the Dutch higher education system.

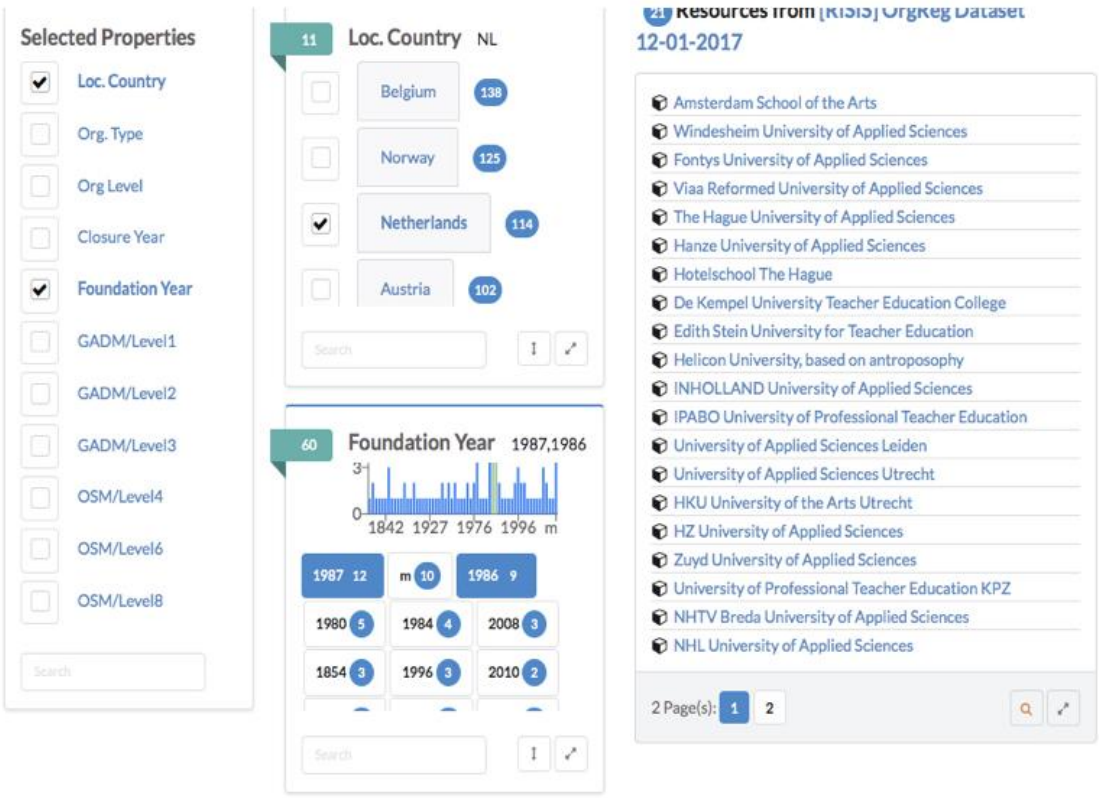


Fig 25. HE institutions Netherlands founded in 1986-1987

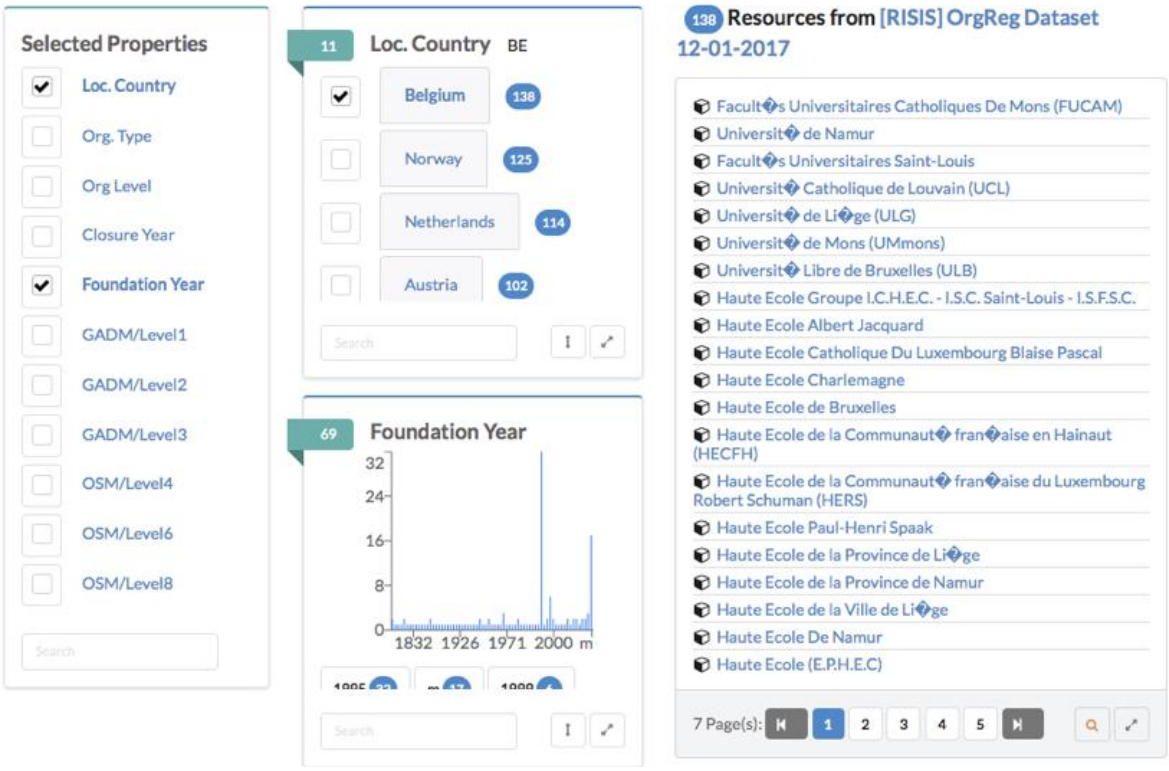


Fig 26. Foundation years HE institutions Belgium

This is a small demonstration of how to use the faceted browser. A follow-up question would be whether this is a typical Dutch phenomenon, or whether similar changes have taken place in other countries.

Belgium could be a second case to inspect, and we do the same steps. Indeed, as the browser shows, also here we find concentrations of foundations of new HE institutions, but now in the year 1995 when 32 new HE institutions were founded in Belgium (figure 26). If we select in the browser the year 1995, we get in the resources list the names of the newly founded institutions (figure 27). We could now further inspect the available information on those institutions, which we haven't done yet. And we do not have prior knowledge on the Belgian system. But inspecting the list of names in the resources table in the figure below, one immediately sees that the changes probably took place in the French speaking part of Belgium, as all names are French language institutions, and not in the Flemish speaking part. Indeed, the two language regions have their own HE system, so this could clearly be the case.

Further data collection is needed to find out what happened in Belgium in the period, and whether it is a similar development as in the Netherlands, but that falls outside the scope of this demonstrator. Here it is sufficient to show how the faceted browser of SMS is useful in such a study.

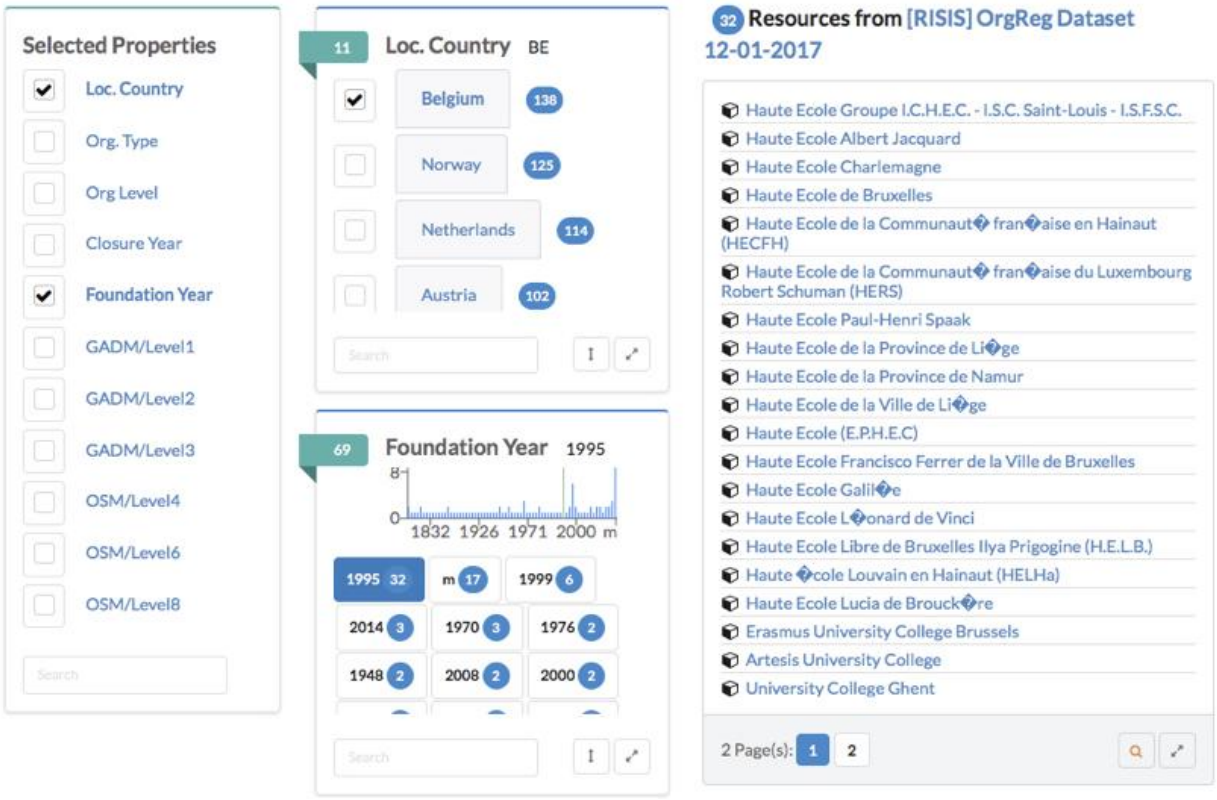


Fig 27. HE institutions Belgium founded in 1995

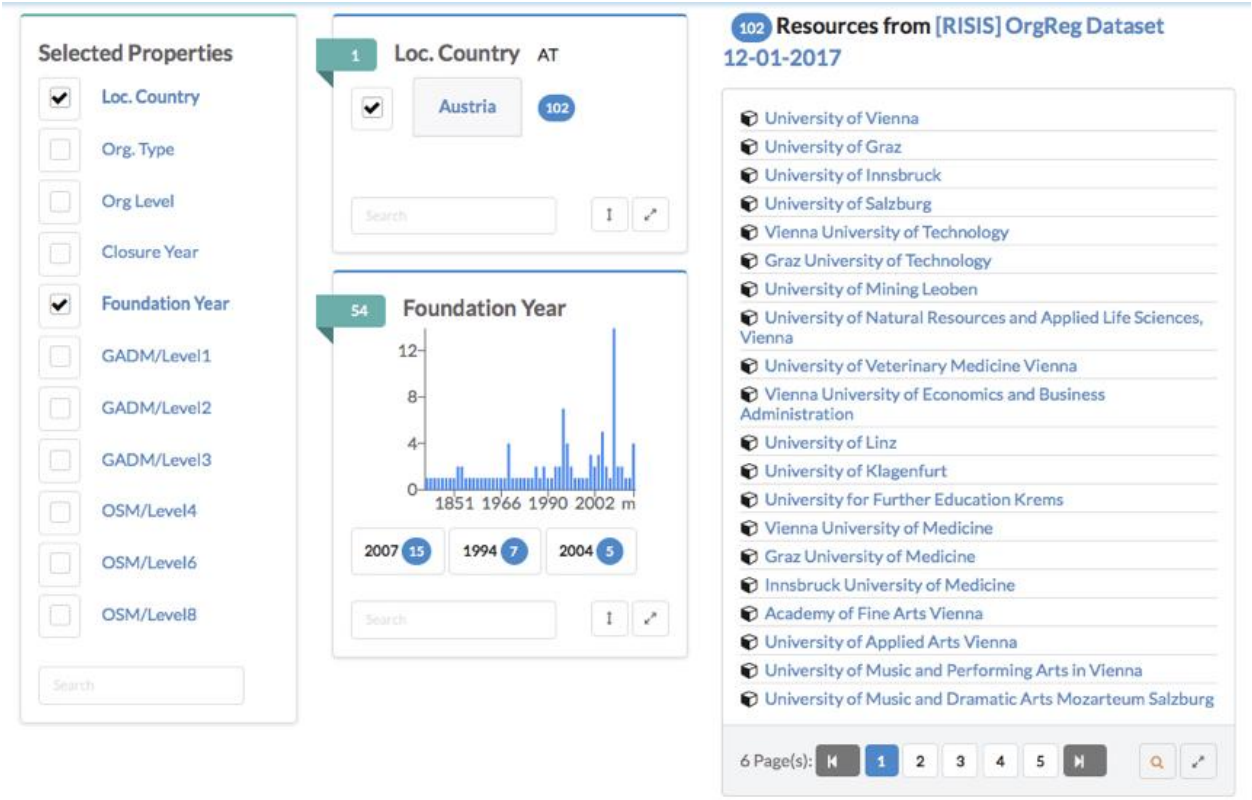


Fig 28. Foundation years HE institutions Austria

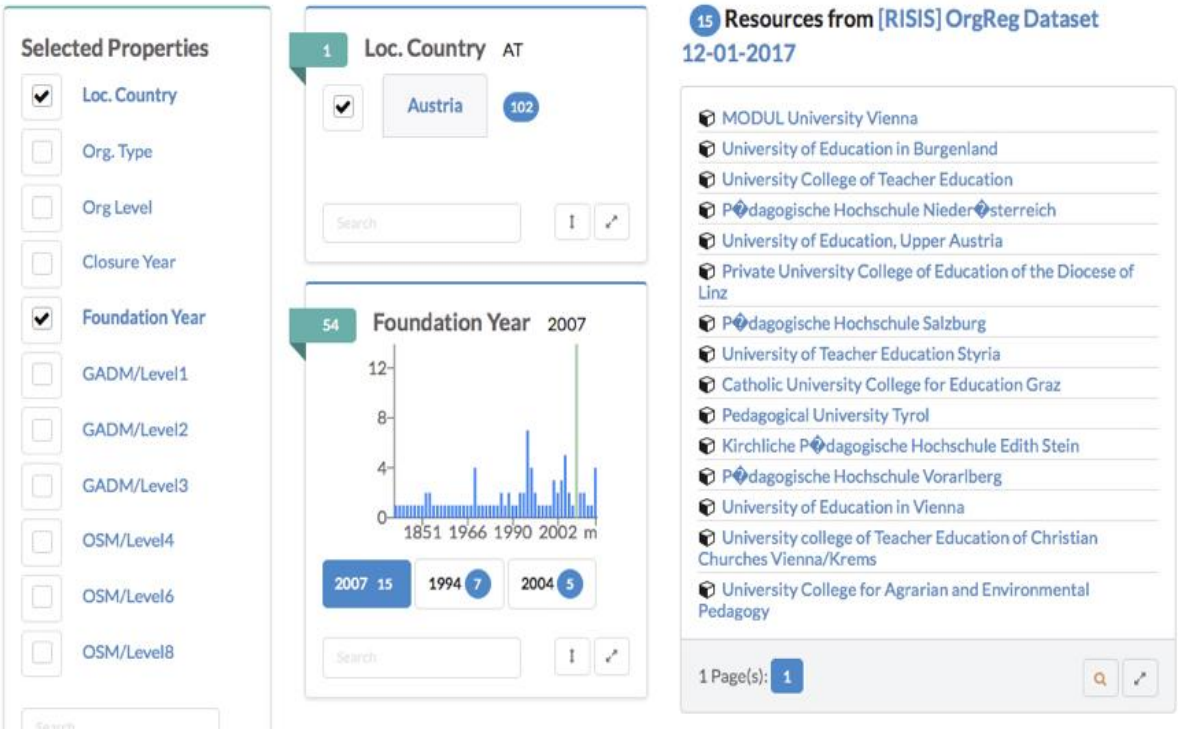


Fig 29. HE institutions Austria - founded in 2007

The third example we give here is Austria (Fig 28 and 29), and indeed also there we detect a concentration of new institutions in 2007 - a decade after the changes in Belgium and two decades after the changes in the Netherlands. Of the total of (now) 102 HE institutions in Austria, fifteen were created in 2007 - again a percentage suggesting some form of structural change. Also in the Austrian case, the browser is helpful. By

selecting the year 2007 in the 'foundation' window, we get in the 'resources' window the list of new institutions.

Even if one is completely unknowledgeable about the Austrian system of Higher Education, the browser tells that the changes have taken place in the sector of teacher education: the newly founded HE institutions are all 'University of Education', 'University College of Teacher Education', and 'Pedagogical University'. Without further investigation, one already can conclude that the changes in the Austrian system are less broad than in the Netherlands or in Belgium, where the changes seem to cover a much larger part of the HE system.

Example 2 Using the open data on organizations for studying links between organizations

A main issue in science and technology studies is the dynamics of collaboration, at the individual level, but also at the level or organizations. As the field is strongly data driven, much of the research operationalized collaboration as 'co-authoring'. Later, studies also used joint projects as a source to study collaboration, which was made possible through the availability of large project databases such as the EC database Cordis (in the SMS platform), and the RISIS dataset EUPRO (partly in the SMS platform). For studying industrial collaboration, often data on joint ventures are collected and used. Here we address the question whether this also can be done for research collaboration. In other words, do public and private research organizations create together new organizations to 'do something together'? Browsing the SMS data store, we do find information about relations between organisations. In the GRID[17] dataset, there are various data on relations between organizations: 'hasChild', 'hasParent' and 'hasRelated' (see figure 30).

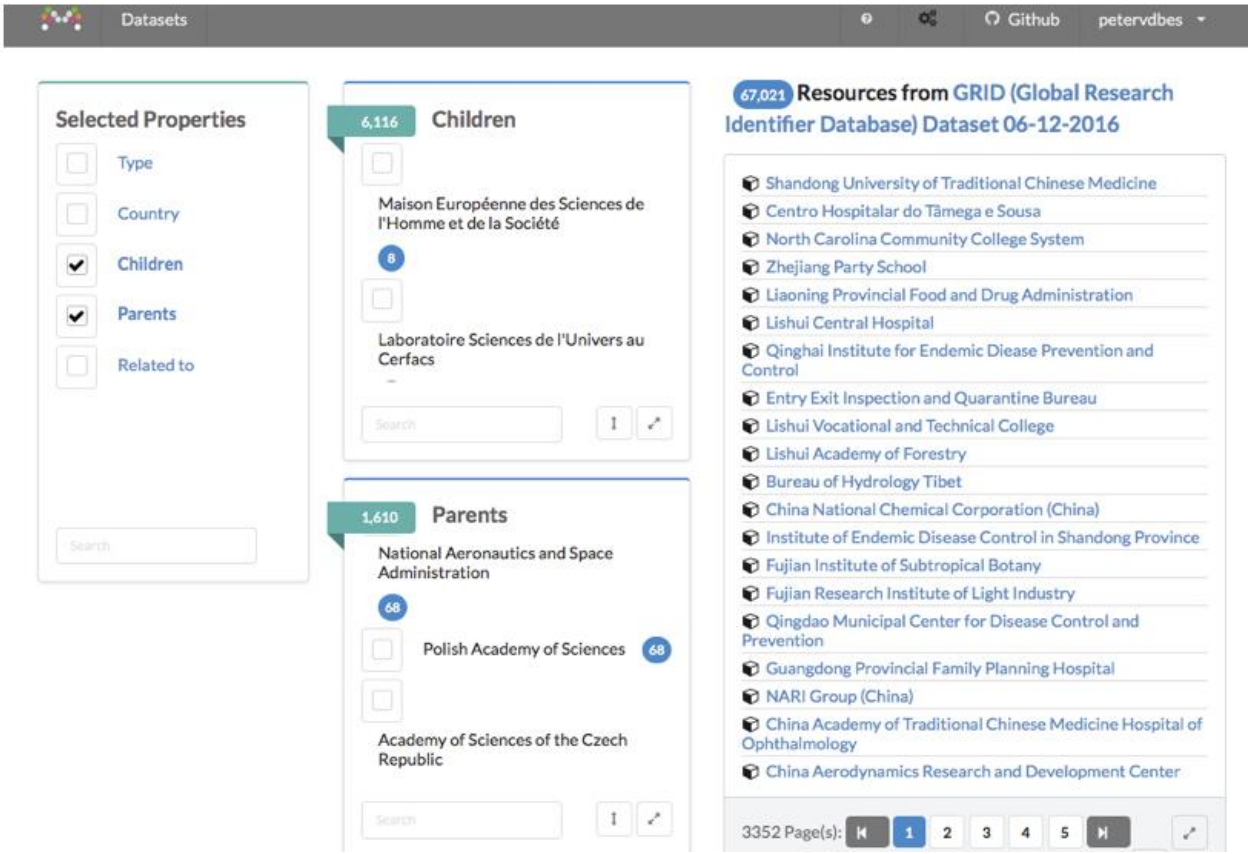


Fig 30. Organizations - 'parents' and 'children'

Using the 'parent-child relations', we now can try to detect the 'joint ventures' in research and higher education. This can be done by selecting properties in the faceted browser, but here we show the result of querying the database. The query asks for all types of organization-pairs, that do have a 'joint venture' relation. In Figure 31, we show the top of the table that the query did produce. We restrict ourselves to joint ventures within countries, as we assume that this is by far the pattern.

[17] https://grid.ac is a reference dataset with organizations that do research, and contains at the moment more than 71,727 organizations worldwide.

Column A gives the country of origin of the organizations. Columns B and C show the sector of origin of the collaborating organizations, and there are several collaboration-types: Education-Government, Education-Education, Education-Facility, Education-Healthcare, government-Government, etc. Column D gives the number of times such a relation-type is in the data, and the last two columns E and F show how many organizations of both types are in the dataset. So in words, row 2 shows that the database includes for France 325 Educational and 168 governmental organization. These span 122 joint ventures.

A	B	C	D	E	F
country	otype	otype2	noWithSharedChildren	totalOType	totalOType2
France	Education	Government	122	325	168
France	Education	Education	121	325	325
France	Education	Facility	84	325	917
France	Government	Education	43	168	325
France	Facility	Education	33	917	325
France	Government	Government	33	168	168
United States	Healthcare	Healthcare	32	2705	2705
United States	Education	Education	28	4101	4101
France	Government	Facility	26	168	917
France	Education	Healthcare	25	325	253
France	Facility	Government	25	917	168
France	Government	Healthcare	18	168	253
France	Healthcare	Government	18	253	168
United States	Education	Government	18	4101	988
France	Education	Other	17	325	170
France	Facility	Facility	17	917	917
France	Healthcare	Education	17	253	325
United Kingdom	Healthcare	Healthcare	17	1254	1254
United Kingdom	Healthcare	Government	15	1254	208
United States	Government	Government	15	988	988

Fig 31. Organizations -joint venture relation by country and type: querying parent-child relations

The table above is sorted descending on column D, se we see here what countries have most joint ventures, and of what type. Obviously, the joint venture model is very popular in France, and therefore we focus on the French joint-venture collaboration network.

As said, it is easy to retrieve the data from the datastore in several formats. So in the next step we retrieve the list of French R&D performing organizations from the dataset, and the list of links between them, where a link is defined as having a child together: 'a joint venture'. These data can then be imported in some analytical tool for network analysis, and here we use Gephi. The next figure shows the result. As we immediately see, the network has a dense core, and a wide periphery (figure 32).

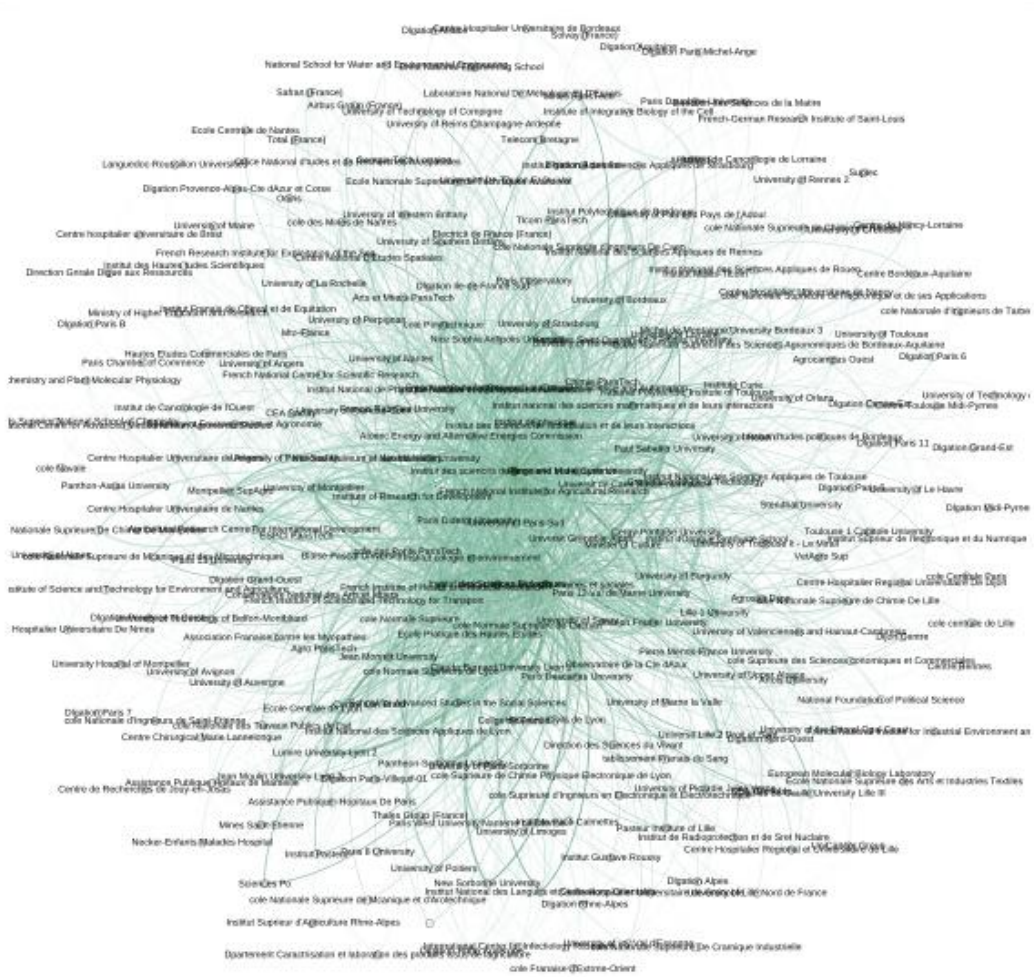


Fig 32. Networks of joint-venture relations: French network. link=shared children

In order to further investigate the network, we calculate a few network characteristics, and one the average degree.The degree of a node is the number of links the node has with other nodes. As 'joint venture' is an undirected link, we do not need to distinguish in-degree and out-degree. The average degree is 20,4 (figure 32) suggesting that jointly creating new organizations is a popular activity in the French system. Or in other words, many research organizations in France seem to be linked to more than one higher level organizations.

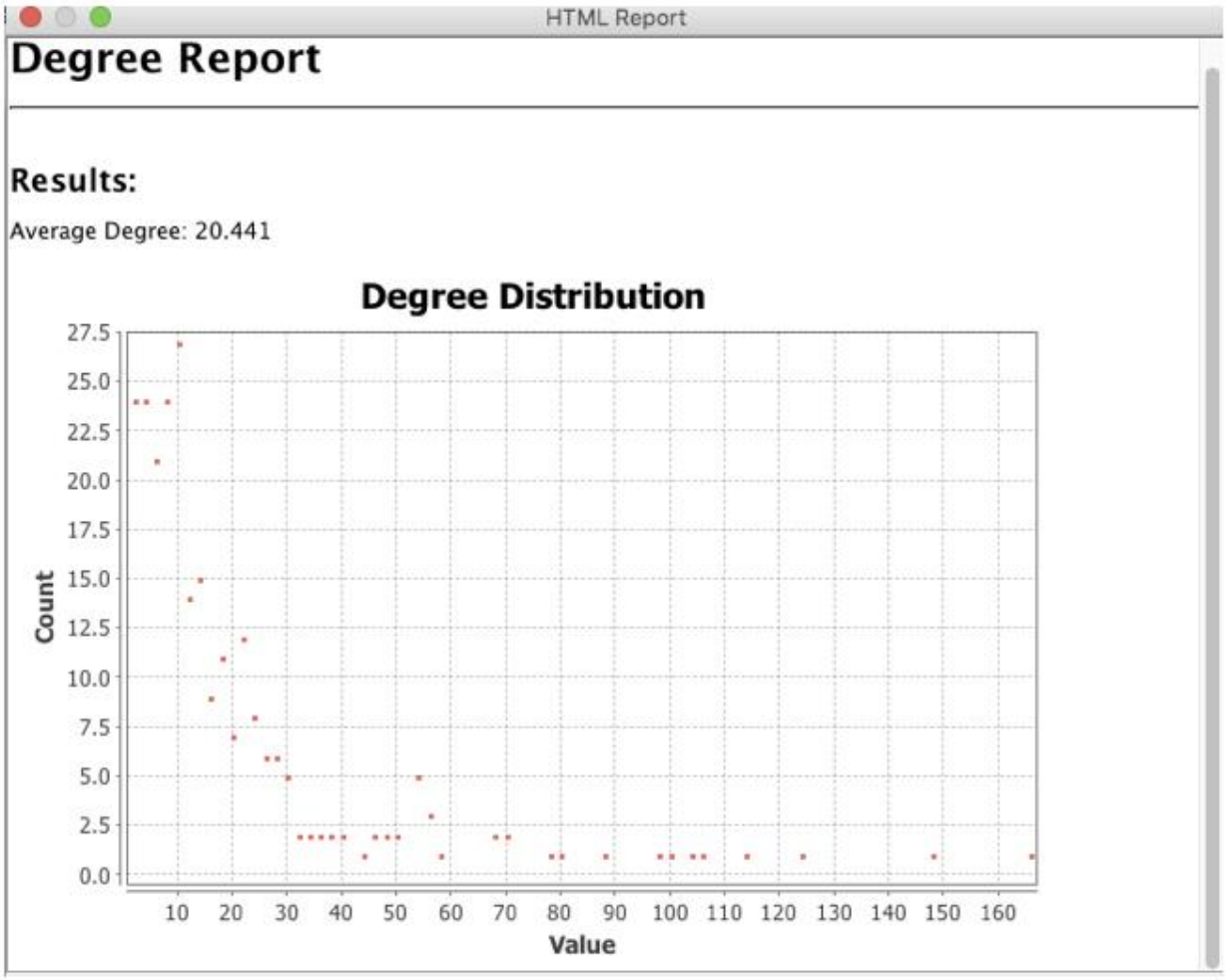


Fig 32. Degree distribution

Nodes Edges Configuration Add node Add edge Search/Replace Import Spreadsheet Export table More actions Filter:

Id	Label	type	l...	long	lat	city	Degree
grid.457016.1	Institut des sciences de l'ingnierie et des systmes	Government		226,403	488,475	Paris	166
grid.457014.3	Institut des sciences humaines et sociales	Government		226,403	488,475	Paris	148
grid.456999.e	Institut des Sciences Biologiques	Facility		237,046	488,179	Ivry-sur-Seine	124
grid.7429.8	French Institute of Health and Medical Research	Government		236,555	488,265	Paris	114
grid.452348.c	Institut National des Sciences de l'Univers	Facility		249,198	488,094	Saint-Maur-des-Fosss	106
grid.457015.2	Institut des sciences de l'information et de leurs interactions	Government		226,403	488,475	Paris	104
grid.414548.8	French National Institute for Agricultural Research	Government		217,077	487,646	Paris	100
grid.457013.4	Institut cologie et environnement	Government		226,403	488,475	Paris	98
grid.457018.f	Institut de physique	Government		226,403	488,475	Paris	88

Fig 33. Organizations by degree

The next indicator is the 'degree distribution', which is shown in the figure below. As often the case, the distribution is rather skewed, and one therefore wonders who these very high linked organizations are. To answer that question, we sort the Gephi data screen on degree, and filter for degree > 80. Figure 32 shows the result, and if one is not familiar with the French system, the next question would be what these 'institutes' in the top of the list actually are.

To answer that question, we use another service of the SMS platform, that is geo-location. The SMS platforms allows the user to find the geographical coordinates for each address, and in fact the platform

does this for the datasets included. As one can see in Figure 33, the OrgRef data are geolocated, and we included the queries in the query. This is now helpful as we can sort the organizations by geocode (figure 34) and this then shows that all these institutes are probably part (divisions?) of CNRS, as they share exactly the same coordinates.

Id	Label	Interval	type	long	lat	city
grid.424348.d	Total (France)		Company	224,321	488,926	Paris
grid.410363.3	Thales Group (France)		Company	224,509	488,922	Paris
grid.4825.b	French Research Institute for Exploitation of the Sea		Facility	225,959	488,228	Issy-les
grid.4444.0	French National Centre for Scientific Research		Government	226,403	488,477	Paris
grid.457013.4	Institut cologie et environnement		Government	226,403	488,475	Paris
grid.457014.3	Institut des sciences humaines et sociales		Government	226,403	488,475	Paris
grid.457015.2	Institut des sciences de l'information et de leurs interactions		Government	226,403	488,475	Paris
grid.457016.1	Institut des sciences de l'ingnierie et des systmes		Government	226,403	488,475	Paris
grid.457017.0	Institut national des sciences mathmatiques et de leurs interactions		Government	226,403	488,475	Paris
grid.457018.f	Institut de physique		Government	226,403	488,475	Paris
grid.462023.1	Direction Gnrale Dlgue aux Ressources		Government	226,404	488,476	Paris

Fig 34. Organizations by geo-coordinates: core of the network = CNRS (geo-location)

One can also try to map geographical and/r functional parts of the network separately, and we use here only the Paris' Higher Education institutions as an example (figure 35).

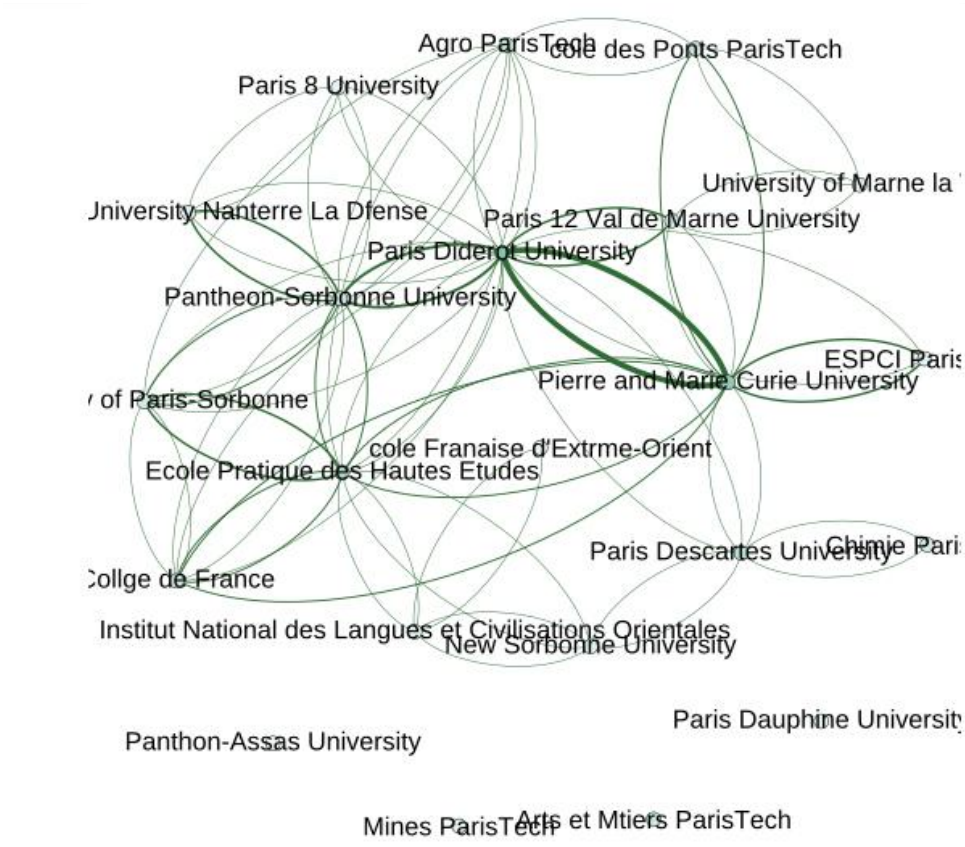


Fig 35. The Paris' universities joint ventures network: link=joint child

Example 3. Using flexible urban areas for studying the localization of innovation

Geography of innovation is another interesting topic. Here we show this might be studied, using the SMS platform. The example we chose is a core element of current science and innovation policy in the Netherlands, where a very large part of public research money is distributed to the so-called 'top sectors',

that is the economic sectors of which Dutch government expects that they will be the core of future economic growth. Money is competitively distributed among consortia that focus on one of the top sectors (such as energy, water, chemistry, life sciences, logistics, etc.), and within the consortia companies have a leading role. Some information about the grated research consortia is available in the RVO project database.

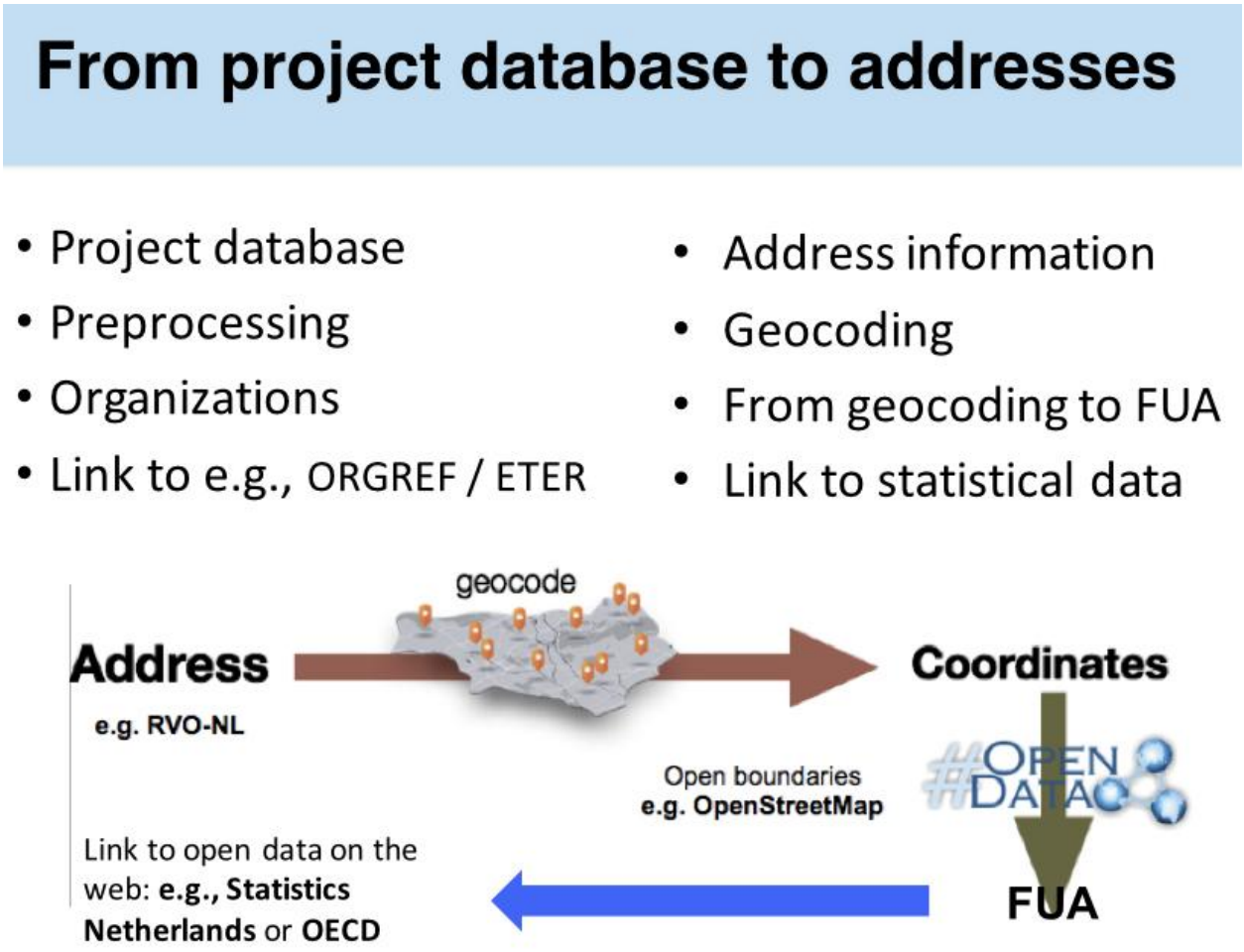


Fig 36. Geography of innovation: data processing

As the 'top sector policy' is considered core in Dutch STI-policy, we are interested in how these research and innovation activities are distributed over the country. How could this be done using the SMS platform? Figure 36 represents the steps, and we will discuss them in some detail.

1. We preprocess the RVO database, and convert the data to RDF.
2. The data are linked to other databases in the SMS platform, which means that we link the names organizations in the RVO database to organizations names in other databases. How this linking is done, and how it can be improved will be described in the next example.
3. Through linking we have more address information, which then is used for geo-coding: finding the coordinates of the organizations involved in the project.
4. The coordinates can be used to find geo-boundaries. A fashionable approach to geo-boundaries are the OECD 'functional urban areas' (FUAs), which the SMS also provides.
5. As many statistical data are provided for regions, we can investigate what characteristics of regions relate to innovation density (in terms of the projects we are investigating.

Figure 37 (left) shows the FUAs in the Netherlands. The map is produced using open data in the SMS platform. Figure 37 (right) shows how the projects are distributed over the FUAs, and the darker the the color the higher the number of projects.

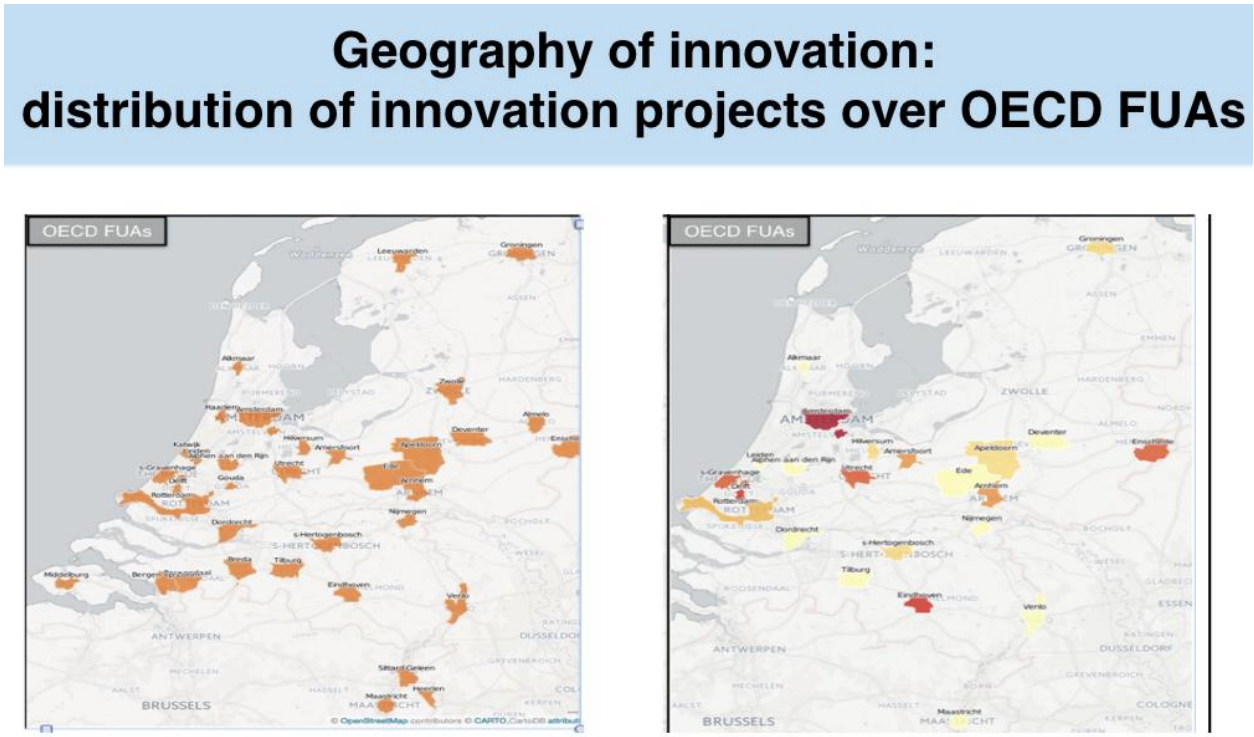


Fig 37. FUAs in the Netherlands (left), and the distribution of projects over the FUAs (right)

Underlying the Functional Urban Areas is an idea of what are meaningful definitions of geographical boundaries. The FUA idea is based on the assumed importance of the distribution of people - population density and commuting patterns. However, a researcher may have good (theoretical) arguments for other geographical delineation. So why wouldn't other characteristics not be more important? the SMS platform has several services to support researchers to use their own definition or geographical areas. These deploy to types of open data. Firstly, we use the availability of statistical data at different levels of aggregation. Second we use the availability of so-called 'shape files' - both available as open data.

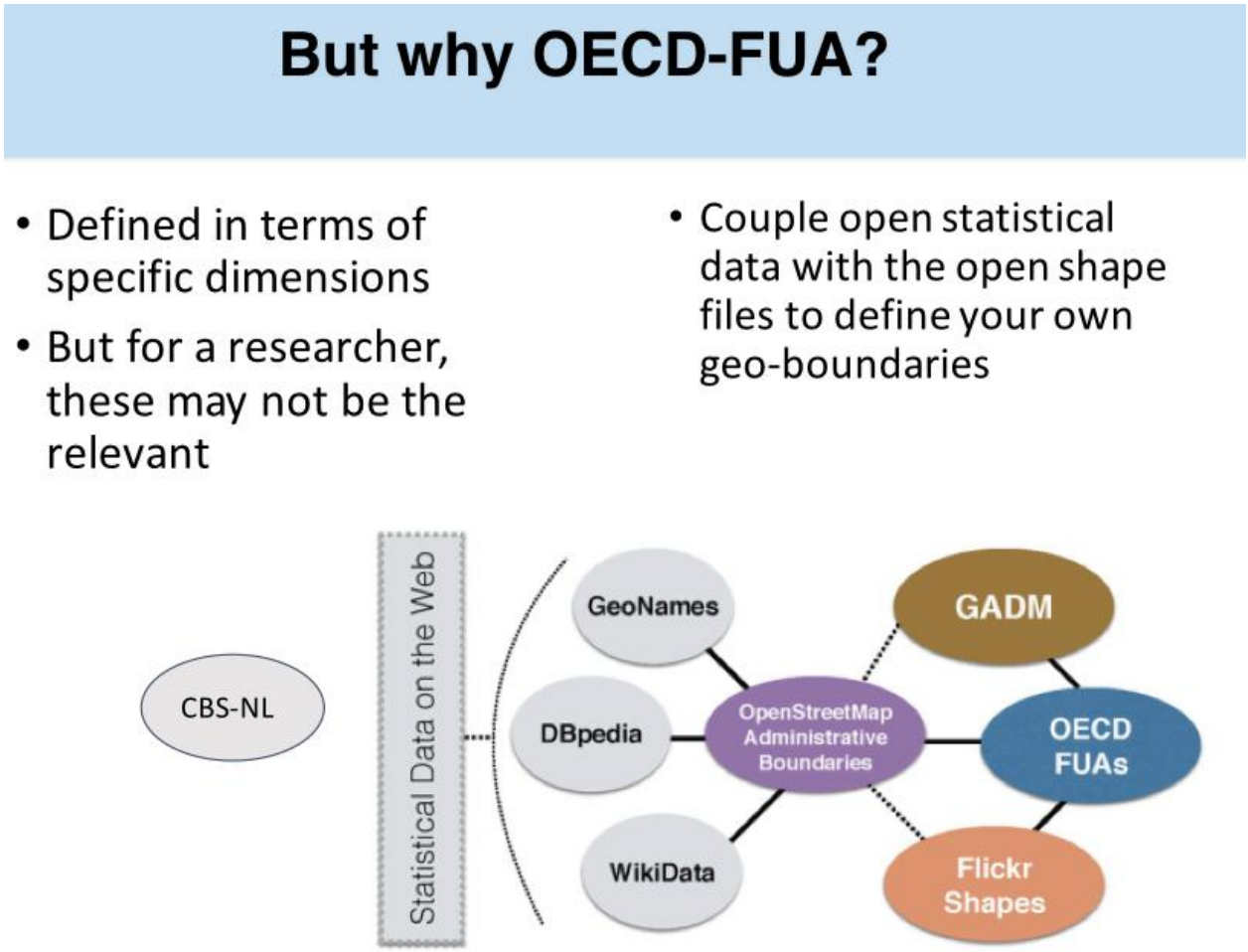


Fig 38. Producing dedicated geo-boundaries

The idea boils down that the researcher defines the property or properties of regions he/she is interested in. An example could be 'population density'. As statistical data are available, the combination of those data with the available geo-boundary data results in a 'population density' geography. Figure 39 (right) shows again the OECD-FUA geography of the Netherlands, whereas figure 39 (left) shows the population density (above a threshold). The two geographies are similar (as population plays an important role in the FUA), but not identical. One may, however, also use a different property such as 'density of companies', or a hybrid definition using as well population density and company density. These definitions result in different geographies, as figure 39 shows.

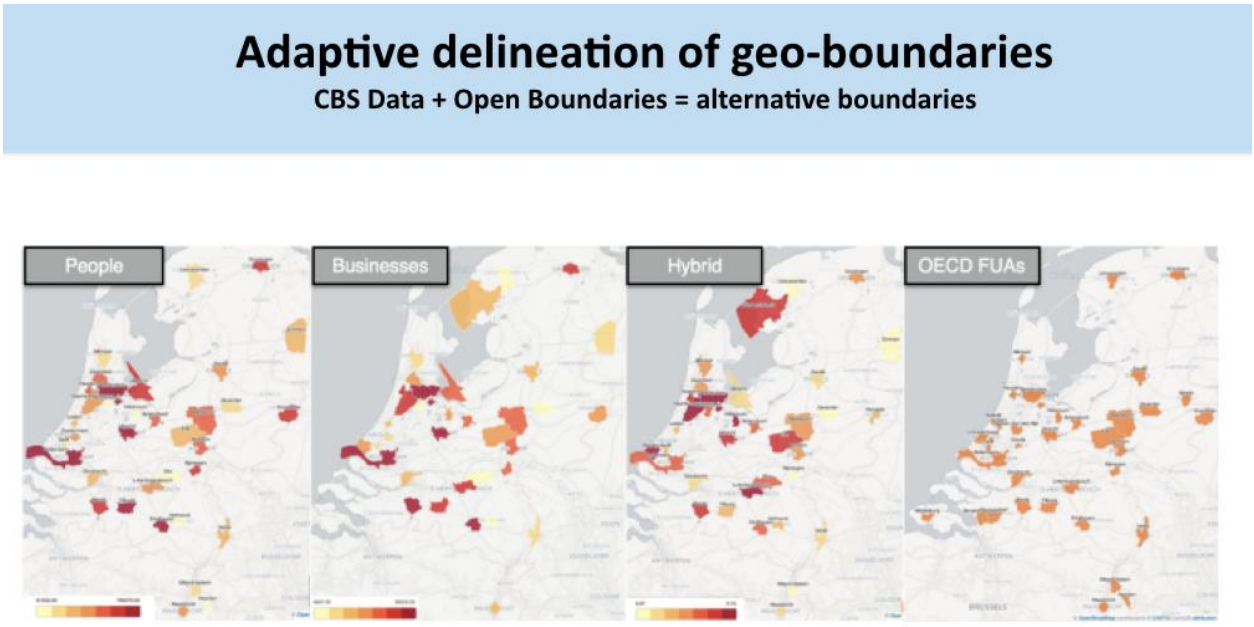


Fig 39. Alternative geo-boundaries

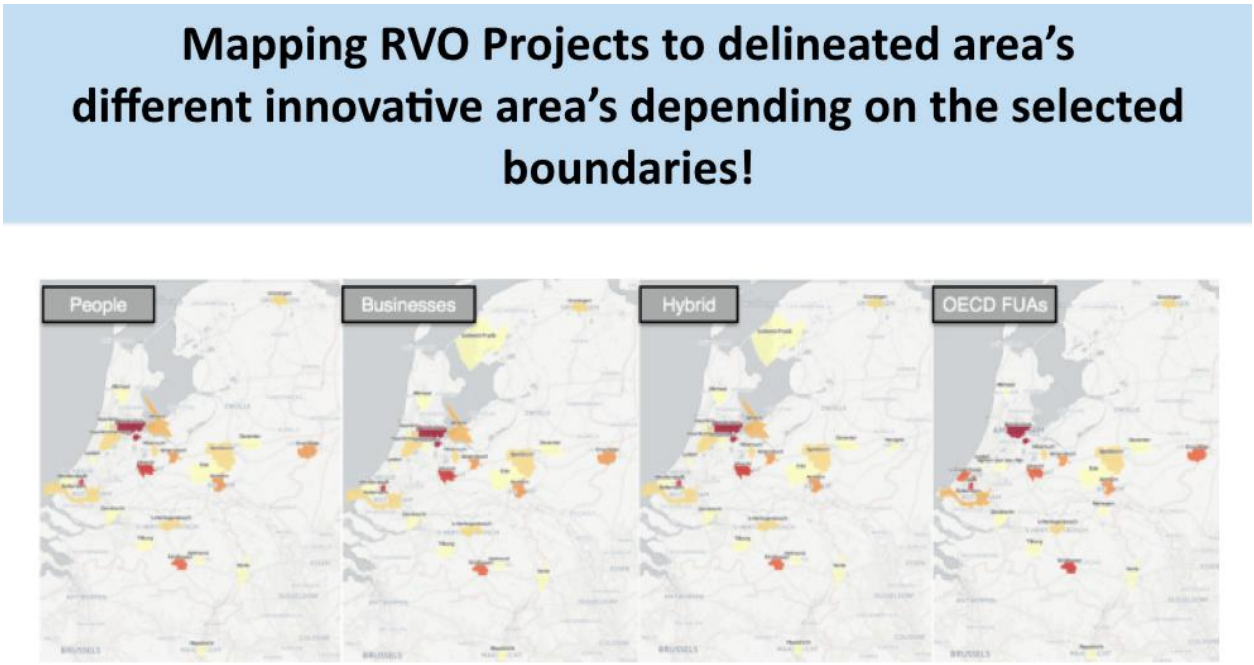


Fig 40: Geography of innovation depends on the selected geo-boundaries

The selection of the definition of geographical boundaries is not neutral when studying the geography of innovation. In figure 40, we show the density of innovation projects (from the 'topsectoren' database) for regions. And obviously, there are regions that do play a role in the innovation projects that are found when using our hybrid definition, but that would have remained invisible using the population density or the FUAs.

The total process for this analysis (see figure 41 for an overview) is rather complex, and requires several computational steps. So doing this may require collaboration between the social scientist who wants to investigate the geography of innovation, and a data scientists to support with the production and retrieval of the required data. But it gives also a flavor of the new possibilities that the SMS platform hopes to provide.

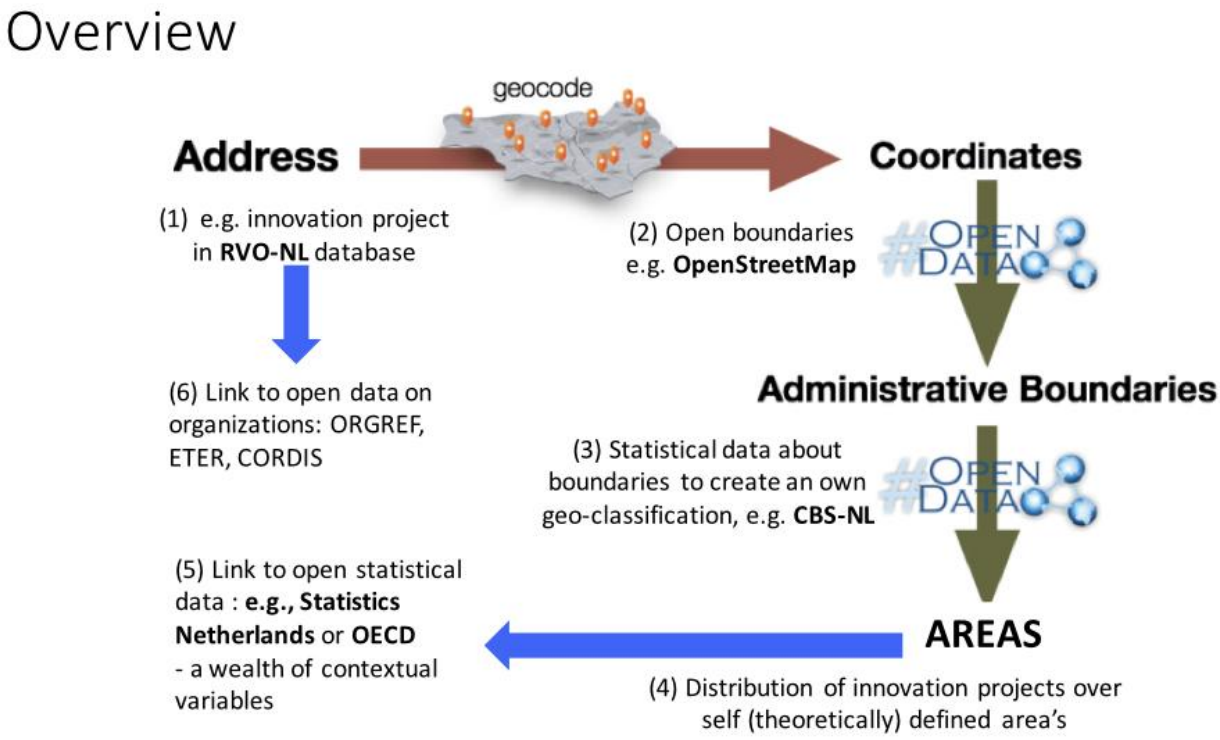


Fig 41. Overview of the data processing

Example 4: Using several sources: does the environment of universities relate to performance?

The last example asks the following question: What characteristics of universities and what characteristics of the environment of universities influence the quality of universities? There are about 2500 higher education institutions in Europe. A few of those are the outstanding - the top ranked - universities, but most of them are much more 'normal'. Why some universities have been the outstanding one's forever is one question, what influences the performance of the large majority is another. One may think of characteristics of the HE institutions, such as size, number of undergraduate and graduate students, student-staff ratio, amount of externally funded research, and so on.

But also contextual variables may have an effect: degree of urbanization, other (higher) education institutions in the vicinity, presence of R&D performing companies, or public research institutes, and other variables representing the social, economic, and demographic characteristics of the region. Why would these factors may be relent. Several theories could be used, but the least one may say is that those social and other factors may affect the attractiveness of the university and the environment for potential students and academic staff. And the more attractive these are, the better staff and students one may be able to select. Another factor may be is that the presence of a variety of research and development and innovation related activities in the vicinity of an HE institution may result in an increase of exchange of ideas, an increase of (interdisciplinary) collaboration, and of funding possibilities.How would be be able to answer these questions? We will focus on the role of the latter factor.

The SMS datastore contains for the moment one dataset with performance data at the university level: the Leiden Ranking. This set contains data for the better but by far not for all HE institutions. Furthermore,the Leiden Ranking only reflects research performance, whereas other rankings also take into account teaching, or external funding (from e.g., industry. In the near future SMS may add some other rankings to increase the scope and the size of the coverage.[18]

The SMS datastore contains several datasets with information about HE institutions, such as ETER, OrgRef, GRID, OrgReg, etcetera. From those we may extract the relevant properties of the HE institutions we are interested in. However, in this example we focus on the contextual factors. How would we retrieve those from the SMS datastore?

The whole process consists of different steps, from linking data, via geo-localization and finding the relevant geo-boundaries, to identifying the other R&D intensive organizations within these geo-boundaries. Then we can measure the number, kind and variety of R&D organizations in the environment of the university as a measure of the quality of the context. Finally we can do some statistics to answer the questions. Does the number, kind and variety of closeby R&D organizations influence the ranking of universities?

[18] A disadvantage of some other rankings is that these are partly reputation based.

Step 1: Linking of the organization names between the relevant datasets, and this is described earlier in this report. In this case, it is about four datasets. After we have done so, we have for all HE institutions a variety of variables, among others the geo-coordinates.

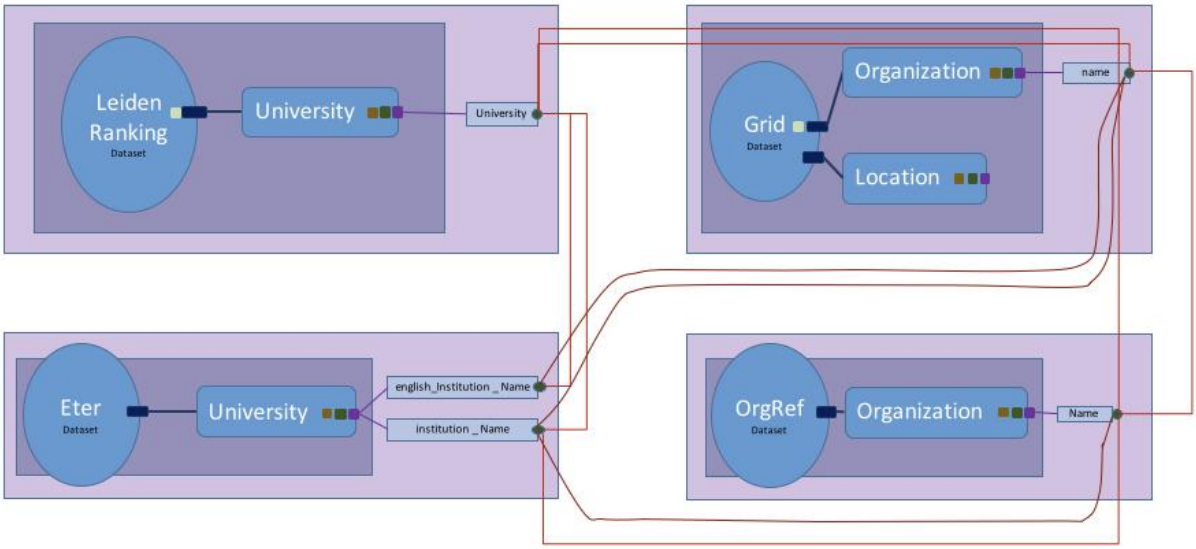


Fig 42. Linking the relevant datasets

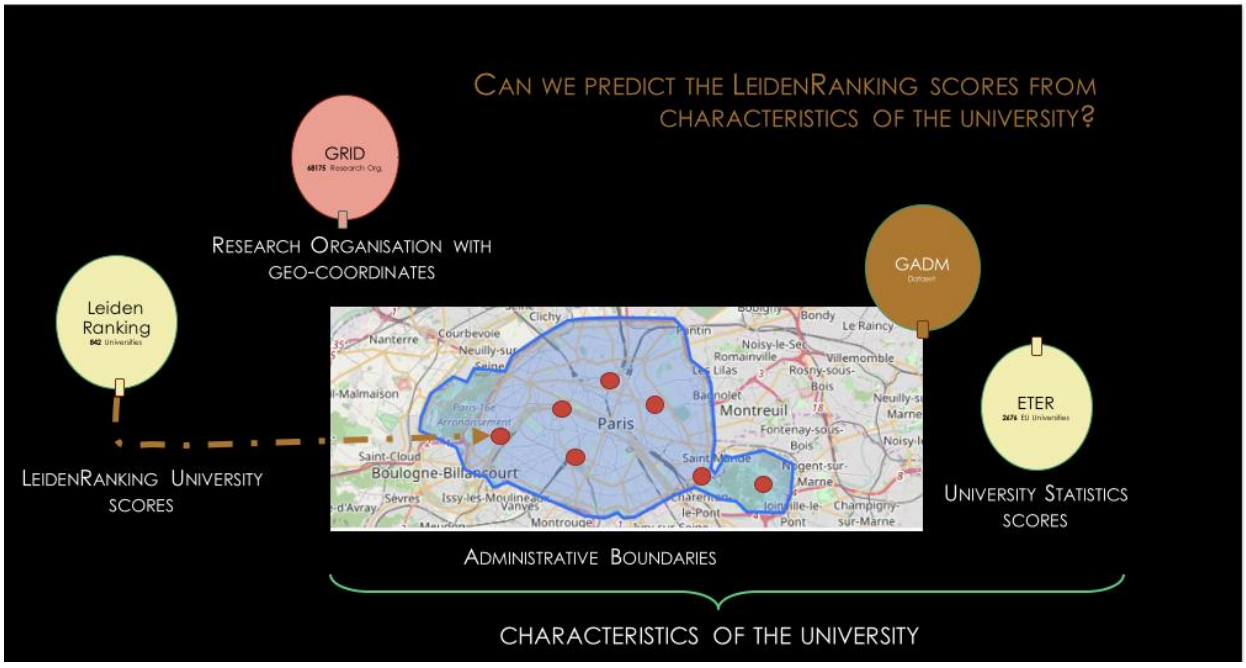


Fig 43. Detecting the other relevant institutions within the environment of an HE institution

Step 2: The geo-coordinates are used to define the boundaries of the environment, and that is needed to find the other R&D intensive within those boundaries.

Step 3: For that we again use the OrgRef dataset, as this contains a huge amount of those organizations, all with their geo-coordinates. For each HE institution, we can now determine which R&D organizations are closeby. As OrgRef also has information about the type of organization, we not only know the number, but also the types, and the variety.

Step 4: These variables, together with the characteristics derived from ETER, can then be used in the explanation of ranking of HE institutions. Figure 44 shows a part of the dataset that can be analyzed in a statistical package like SPPS SAS, or R. The 'english_name', 'country', 'category', 'total_expenditure', 'third_party _funding' and 'Academic staff size' are all retrieved from ETER. The performance score 'PP_top10' comes from the Leiden ranking, the 'longitude and 'latitude' come from GRID, and the 'geo-boundary' is produced in the SMS platform. The geo-boundary and GRID are used to calculate the 'Number of R&D intensive organizations'.

To what extent these variables indeed predict the ranking is to be answered - but the correlation between the two yellow columns (with the Netherlands universities only) is 0.58.

english_name	Country	Category	latitude	longitude	geoboundary	number R&D orgs	totaExpenditureEURO	thirdPartyFundingEURO	totalAcademicStaffFTE	PP_top10
University of Amsterdam	Netherl	univers	52,368,941	489,127	<http://geo.risis.eu/gadm/158-9-2	79	596943000	95795000	2,530	17.10%
VU University Amsterdam	Netherl	univers	5,233,356	4,864,845	<http://geo.risis.eu/gadm/158-9-2	79	451900000	91000000	2,205	16.50%
Utrecht University	Netherl	univers	52,084,918	517,383	<http://geo.risis.eu/gadm/158-11	56	756409000	223587000	2,694	17.50%
Leiden University	Netherl	univers	52,156,535	4,486,543	<http://geo.risis.eu/gadm/158-14	30	488500000	163600000	1,938	13.60%
Erasmus University Rotterda	Netherl	univers	51,919,779	4,524,159	<http://geo.risis.eu/gadm/158-14	25	518800000	156800000	1,178	17.70%
Eindhoven University of Tech	Netherl	univers	51,447,954	5,485,308	<http://geo.risis.eu/gadm/158-8-2	17	314600000	98600000	1,792	13.40%
Delft University of Technolog	Netherl	univers	52,002,726	4,375,193	<http://geo.risis.eu/gadm/158-14	14	524441000	143345000	1,962	15.00%
Radboud University Nijmege	Netherl	univers	51,819,359	5,857,048	<http://geo.risis.eu/gadm/158-4-9	13	500250000	149617000	1,852	16.30%
University of Groningen	Netherl	univers	53,219,235	656,373	<http://geo.risis.eu/gadm/158-5-1	12	595477100	151612500	2,130	15.70%
Maastricht University	Netherl	univers	50,846,816	5,686,782	<http://geo.risis.eu/gadm/158-7-1	11	346946000	85802000	1,818	15.30%
University of Twente	Netherl	univers	5,223,877	6,850,542	<http://geo.risis.eu/gadm/158-10	11	310800000	83400000	1,573	13.90%
Tilburg University	Netherl	univers	51,563,139	5,040,706	<http://geo.risis.eu/gadm/158-8-2	6	201054151	53156593	899	12.40%

Fig 44. Part of the resulting dataset (Dutch universites only, and a few of the variables)