Towards an open infrastructure for Science, Technology and Innovation data[1], [2]

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

Al Idrissou[#], Antonis Loizou*, Stefan Schlobach*, Frank van Harmelen*

Vrije Universiteit Amsterdam, Network Institute,
[#] Department of Organization Sciences, * Department of Computer Science

[p.a.a.vanden.besselaar][a.khalili][ka.de.graaf][o.a.k.idrissou]
[a.loizou][k.s.schlobach][frank.van.harmelen]@vu.nl

Abstract

In this paper we describe the SMS 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. An example shows how the platform enables data integration in practice. In another example we illustrate how the platform can create and adapt alternatives to the OECD Functional Urban Areas (FUAs) by integrating data from multiple up-to-date open data sources.

1. Introduction

In this paper we describe the Semantically Mapping Science (SMS) data integration platform for STI data using semantic web technology and focusing on (but not restricted to) linked open data (Beek et al 2016). Figure 1 shows the basic architecture we will describe in more detail in this paper.

Why is such infrastructure needed? 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 infrastructure is the conversion of different datasets in the same open format: from tabular data, text data and web data to RDF (Resource Description Framework) data (Beek et al 2016).

[1] The work underlying this paper is supported by EC grant 313082, the RISIS project: Research Infrastructure for Science and Innovation Studies.

[2] Paper presented at the 3rd OECD Blue Sky Conference - Ghent, Belgium, Sept 19-21 2016

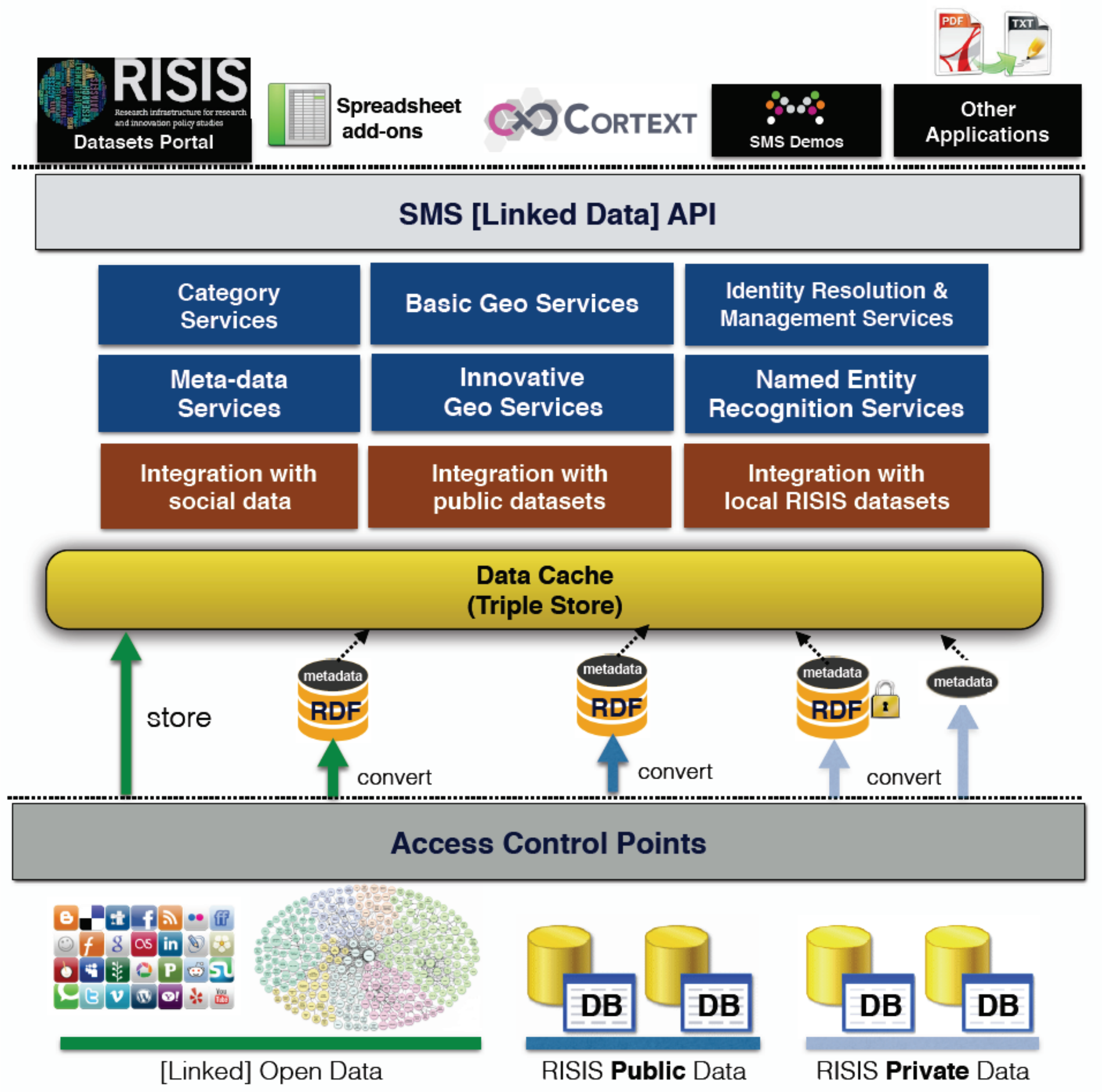


Figure 1: the basic SMS architecture

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. The SMS platform is partly implemented; we aim at providing a complete beta version later this year, as part of the RISIS science and innovation studies data infrastructure (www.risis.eu). Below we first describe the architecture and then the different functions that the SMS platform offers. From the user perspective the front end of the system is important: the interfaces that enable to retrieve the data, the services that enrich and interlink data, and the applications that analyse and visualize the data. The integration and refinement of linked open data source is important from the perspective of developers, knowledge engineers, and policy makers. The dataset-linking and geo-location functions are both illustrated by a case.

2. Architecture of the SMS platform

As shown in Figure 1, 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.

Access Control Points

We are dealing with three types of datasets within the RISIS project:

- *Private (RISIS) Datasets.*
  These datasets contain private, confidential, and/or sensitive data or data for which a specific permission or subscription is required.
- *Public RISIS Datasets*.
  These datasets can be exposed openly to users for research purposes.
- *Public Open Data.*
  These are all available public data on the Web, which can be integrated with current RISIS datasets for the sake of disambiguation or enrichment.

Access Control Points are gatekeepers between the SMS-RISIS platform and the available datasets. They provide two main features:

- *How to get access to data?*
  The data on RISIS is stored in different heterogeneous formats (e.g. Excel, CSV (Comma-Separated Values), Microsoft Access, SQL (Structured Query Language), RDF, etc) and by different storage systems. Access Control Points will provide standard interfaces which will reduce technical difficulties of accessing data.
- *How to control and manage access to data?*
  There are different levels of accessing data. One can only access specific entity types or attributes of a certain dataset. Access Control Points will provide a mechanism to coordinate access to data based on the user role and the dataset owner's requirements.

Data Conversion

RDF (as Linked Data standard) is the main data model used within the SMS platform. In order to process data in SMS, all the incoming data need to be converted to RDF. This is done by incorporating existing RDB-to-RDF (Relational DataBase-to-RDF) tools or by developing custom mappings and conversion mechanisms.

Data Storage

Once the data is converted to RDF, SMS platform will use Triple stores as efficient ways of storing Linked Data in the platform. The graph-based storage will provide the context for data interlinking and querying.

Data Integration

Within the SMS platform, different disambiguation mechanisms will be applied to the stored RDF data to create links between different datasets. The links will be added back to the storage system together with their specific justification.

Web Services

The interlinked data provides a basis for running data integration queries. These queries are written in SPARQL (SPARQL Protocol and RDF Query Language). In order to simplify access to the result of data integration, a set of services with standard interfaces are provided.

Applications

In order to demonstrate the functionality of SMS Linked Data Web services in specific domains, a set of applications are built on top of SMS services. In the next sections of this paper, we will describe some of these applications.

3. Functions of SMS platform

Preprocessing heterogeneous data

Figure 2 shows how the platform supports pre-processing and conversion of data into the RDF standard for linked open data. For example, PDF files can be converted into TXT (text) files, and through *Named Entity Recognition* web service relevant entities like people, organizations, countries, etc. are identified.

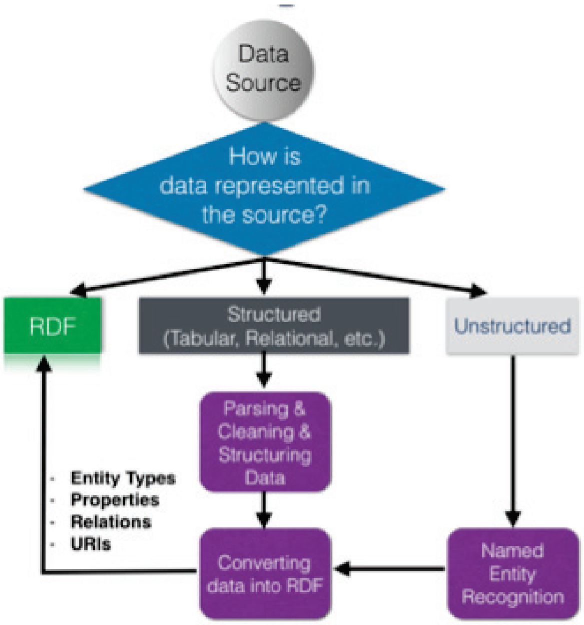


Figure 2: From heterogeneous data to RDF

Additional text processing (e.g., term extraction) may identify attributes. Structured data (e.g., Microsoft Excel files) are parsed, cleaned, and then converted into RDF. A concrete example is recognizing research institutions and universities in a researcher’s CV (Curriculum vitae), using name recognition by linking the CV to databases with background knowledge such as DBpedia. The resulting data are then converted into RDF.

Linking datasets

The next step is *linking* the data. If entity *identifiers* are available, the linking is easy. If not, a variety of techniques can be used, from (fuzzy) *string matching* to *deploying attributes* available in the different databases. If names occur in different languages, resources like DBpedia can be used to match. If two entities have different names, but similar other characteristics (such as geo-location), they may be in fact the same entity.

However, whether entities are considered the same, depends on the perspective: sometimes two organizations (e.g. departments) can be the same – because they are parts of the same organization (university). But if one wants to compare departments, this is not anymore the case: the departments then need to be considered as different organizations. We are currently experimenting with a series of datasets on research organizations, in order to compile basic reference sets of research organizations. This is done through interlinking different datasets through knowledge resources on the web (Figure 3).

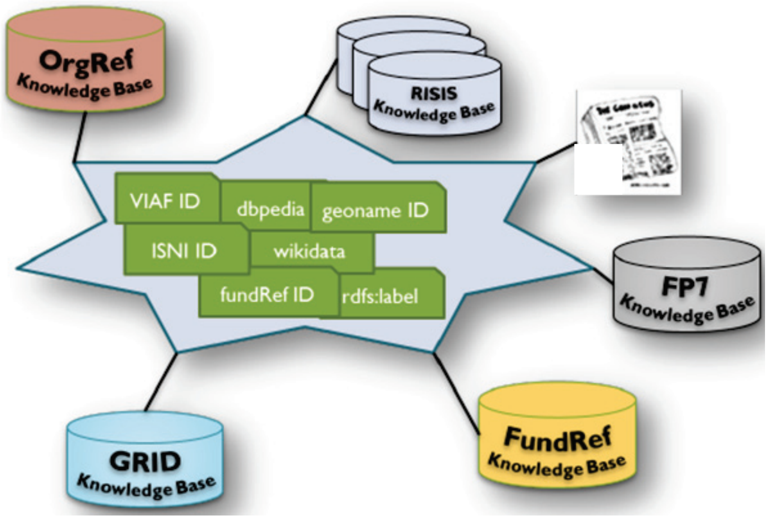


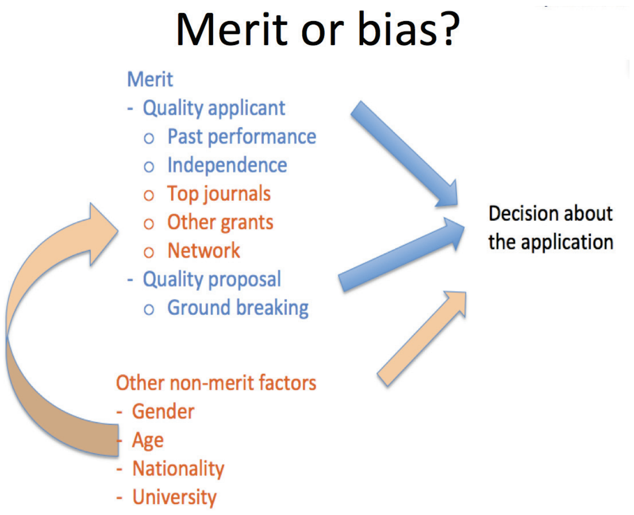
Figure 3: Linking data through web knowledge resources.

Linking data in practice - An example

In a project we investigated whether gender of applicants influences the grant decision. In order to answer that question, one needs a multitude of variables that may influence the grant decision - as suggested by various theories: variables representing merit (such as scholarly performance), but also variables bring in bias, such as personal characteristics (gender, age, nationality, university of PhD degree). And some in-between variables, such as the quality of the applicant's network. We use the SMS platform for preprocessing, for converting into RDF, for entity recognition and for linking. These variables come from a variety of data sources:

- Bibliometric performance scores: Web of Science (TXT)
- Quality of the applicant's network:
  - Organizations mentioned in the CV (PDF)
  - Ranking of those organizations from Leiden Ranking (Excel)

- Earlier grants: from CV (PDF)
- Host institution from admin file (Excel)
- Ranking of host institution from administrative file (Excel)
- Personal characteristics from admin file (Excel)
- Evaluation language: Term extraction from review forms (PDF)



Merit or bias?

Merit
- Quality applicant
  - Past performance
  - Independence
  - Top journals
  - Other grants
  - Network
- Quality proposal
  - Ground breaking

Other non-merit factors
- Gender
- Age
- Nationality
- University

Decision about the application

Geo-services

An interesting possibility is *linking through geo-location*: if two entities have the same geo-location, they may be related (or identical). Geo-locating has an additional advantage, as it is also an instrument to enrich data: many other (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 platform provides a variety of *geo-services.* The geo-services system is based on a series of open geo-resources, such as GADM[3], OpenStreetMap[4] and Flickr geotagged data[5]. By integrating these geo-resources, the service can give for an entity's address the geo-location up to 11 different levels (Figure 4).

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 5, in the top left part of the screen the address "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.

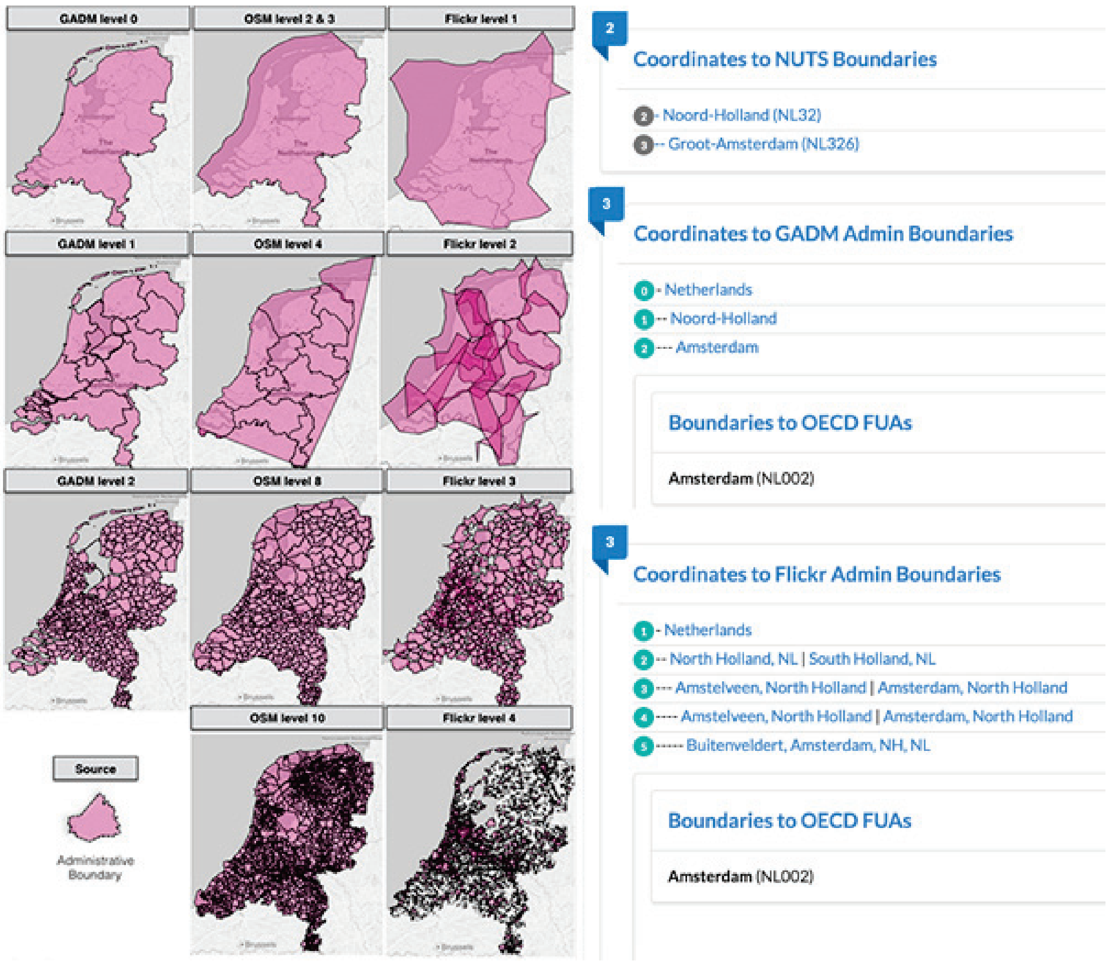
Figure 5 shows the various administrative boundaries for the geocoded address. in this case, level 8 represents LAU 2 (Local Administrative Unit). The platform can be used to do this for large(r) amounts of addresses, and the output then is not on the screen, but in a tabular form.

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

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

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

For this purpose, the SMS platform provides a *spreadsheet add-on* where users can enter a (long) list of addresses to be geocoded. In the future we aim at adding different distance concepts, such as travel distance (time, frequency, price, etc.) as part of the innovative SMS geo-services.



Coordinates to NUTS Boundaries

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

Coordinates to GADM Admin Boundaries

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

Boundaries to OECD FUAs

Amsterdam (NL002)

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

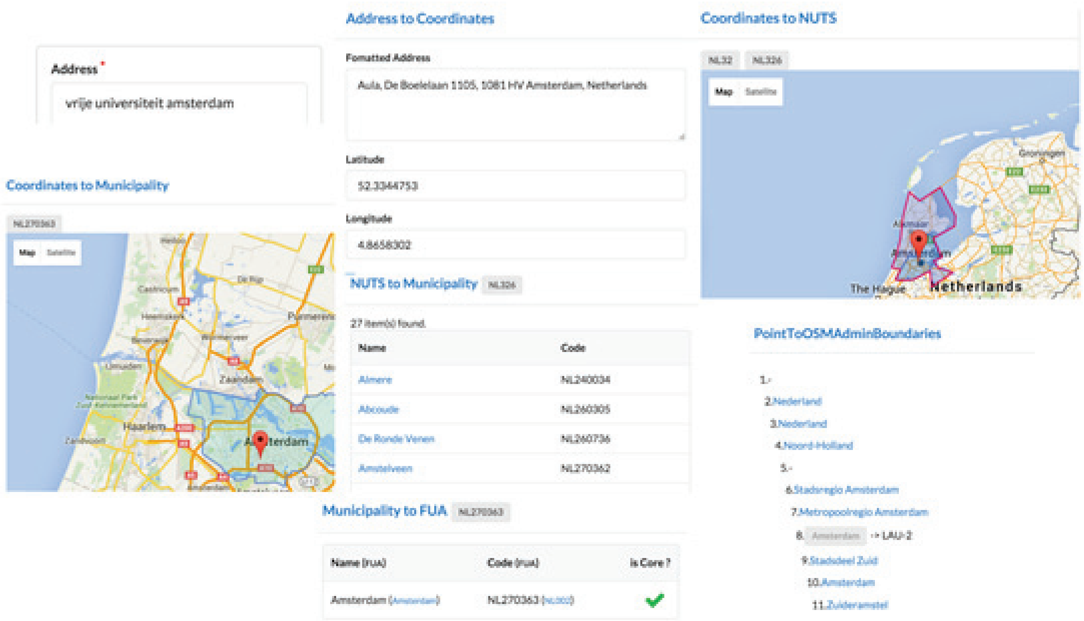
Boundaries to OECD FUAs

Amsterdam (NL002)

Figure 4: Mapping administrative boundaries using open geo-resources

The advantage of this service is twofold. Firstly, through location of an entity (e.g., a university) within a *given administrative boundary* (e.g., a municipality), one can link the entity to other (e.g., statistical) data available at that geographical level. For example, National Statistical Agencies like Statistics Netherlands (CBS)[6], and international organizations like OECD and Eurostat, publish many socio-economic, population, transport, environmental, and other data at several spatial levels such as regions, provinces, municipalities and neighborhoods.

[6] https://www.cbs.nl



Address *

vrije universiteit amsterdam

Address to Coordinates

Fomatted Address

Aula, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands

Latitude

52.3344753

Longitude

4.8658302

Coordinates to NUTS

NL32 NL326

Coordinates to Municipality

NL270363

NUTS to Municipality NL326

27 item(s) found.

Name	Code
Almere	NL240034
Abcoude	NL260305
De Ronde Venen	NL260736
Amstelveen	NL270362

Municipality to FUA NL270363

Name (FUA)	Code (FUA)	Is Core ?
Amsterdam (Amsterdam)	NL270363 (NL002)	✓

PointToOSMAdminBoundaries

1.-
2.Nederland
3.Nederland
4.Noord-Holland
5.-
6.Stadsregio Amsterdam
7.Metropoolregio Amsterdam
8. Amsterdam -> LAU-2
9.Stadsdeel Zuid
10.Amsterdam
11.Zuideramstel

Figure 5: Geo-locating services

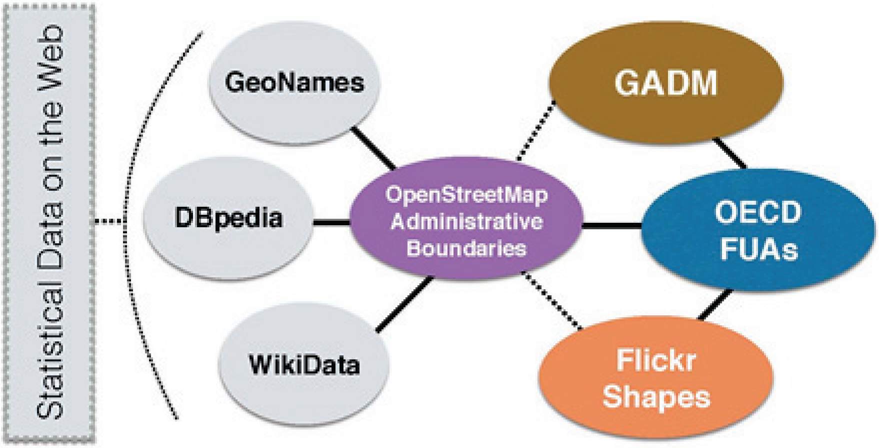

Statistical Data on the Web
GeoNames
DBpedia
WikiData
OpenStreetMap Administrative Boundaries
GADM
OECD FUAs
Flickr Shapes

Figure 6: Interlinking administrative boundaries to existing Linked Open datasets

Secondly, interlinking these statistical data with geographical data from administrative boundaries, *alternative geographical schemas* similar to OECD's Functional Urban Areas can be provided. For example, indicators such as population, living costs, income, facilities, power consumption, can be taken into account for defining the (administrative) geographical boundaries. This enables a researcher to create tailored geographical boundaries defined in terms of the characteristics he/she considers relevant. As shown in Figure 5 for the Netherlands, different levels of boundaries are extracted which can be mapped to statistical data provided for those boundaries. Figure 6 shows the links between the extracted boundaries and open datasets such as DBpedia, WikiData and GeoNames.

Data enrichment through geo-location: An example

As explained, one of the objectives is to integrate different science and technology datasets with open data, using geographical mapping. How would that look like? As a use case in this direction, we investigate the effects of socio-economic and structural properties of urban areas on the level of innovative activities, as stimulated by recent research and innovation policies in the Netherlands. As the focus of this paper is not the use case as such, but the SMS-platform, we only describe it briefly.

The new policy in the Netherlands is oriented at the 'top sectors' of the economy, which were selected after consultation of policy makers, representatives of the research system, and entrepreneurs in the country. After selecting these 'top sectors', a large part of public research funding was devoted to this new policy. Consortia can apply for funding, and they should exist of companies and research organizations (such as universities) with a company as main applicant. Because of this context, the funded projects can be considered as a useful representation of RTD (Research Technology Development) collaboration for innovation. In this case we are interested in the geographical properties of these consortia. In order to investigate this, one needs data about the projects and data about the characteristics of the relevant geographical units. These data are available as open data. In this case, the following open datasets are deployed:

- RVO dataset[7] providing a list of research and innovation projects that have received subsidies and financial support from the Netherlands Enterprise Agency (Rijksdienst voor Ondernemend Nederland). Projects information includes companies and research institutes that are collaborating in the project, but also the geographical coordinates of the project main applicant.
- CBS dataset[8] published by the Statistics Netherlands provides different types of statistical information on dimensions such as labour and income, economy, society and regional aspects of municipalities and regions in the Netherlands.

As one does not know *ex ante* what the level of geographical organization of the consortia is, we need to define the ‘geographical containers’ in different ways. That would enable us to find out at what geo-level these consortia are in fact organized. Then we can identify the relevant characteristics of these geographical 'containers' of the projects. As an example, we first calculated different sets of Urban Areas based on different statistics provided by the CBS dataset and different levels of open administrative boundaries. Figure 7 shows the delineation of these Urban Areas through (i) the population size, (ii) the number business establishment and through (iii) combinations of these two indicators at the municipal level. The geo-location service enables the user to put different weights on different indicators when delineating the boundaries, which could be used for an analysis of the role of specific factors. Boundaries typically differ when defined by different characteristics. When compared with the OECD FUAs[9], the adaptive Urban Areas take into account additional

[7] http://www.rvo.nl/open-data-van-rvonl

[8] https://www.cbs.nl/en-gb/our-services/open-data

[9] We used the shapefile from the Eurostat used in the large study named urban audit: http://ec.europa.eu/eurostat/web/cities/data/database

regions (administrative boundaries). Figure 7 show the different maps of urban areas that result from different definitions. The darker the area is, the higher the score on the variables used to define the Urban Areas.

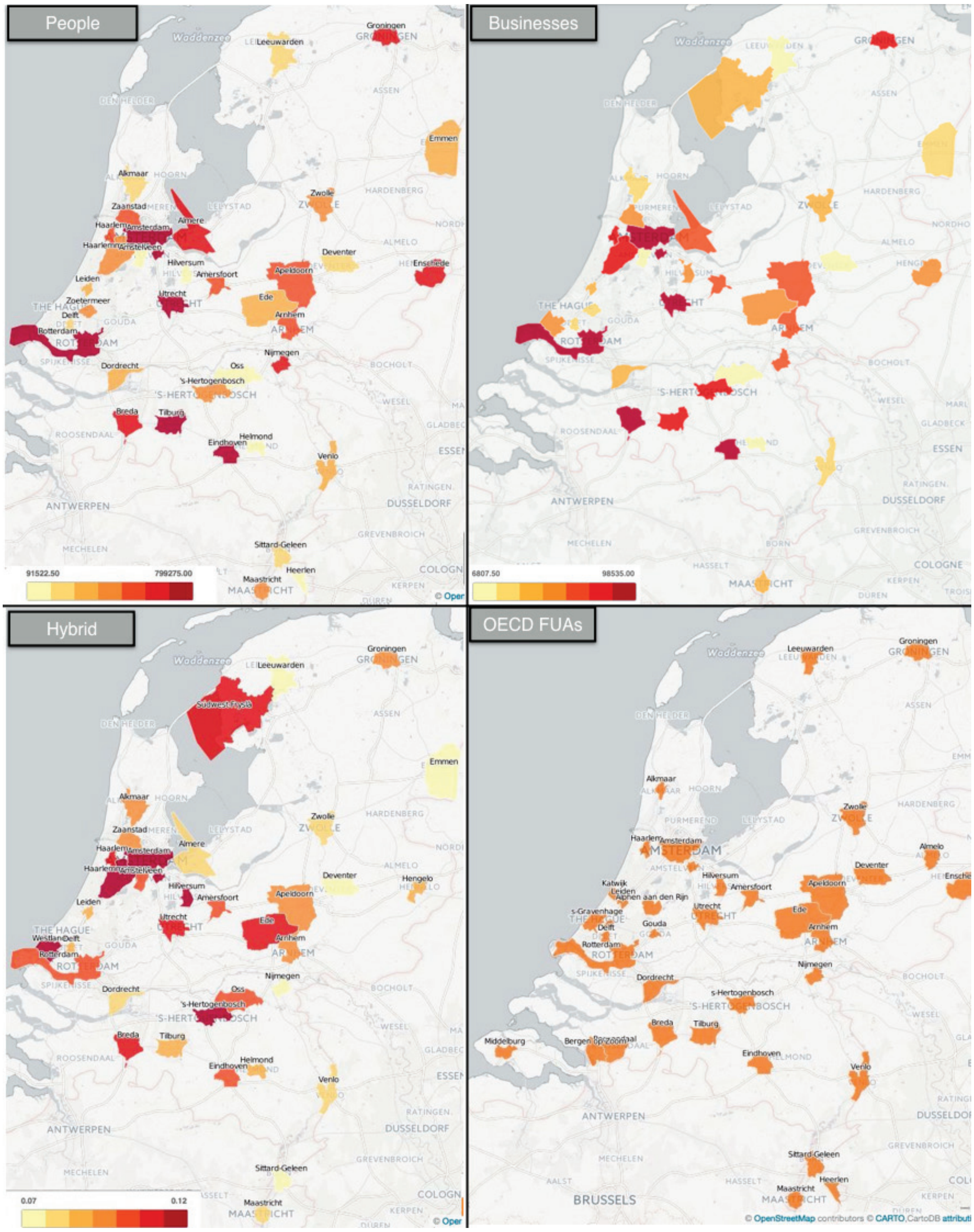


Figure 7. An example of the adaptive delineation of FUAs for the Netherlands based on the open statistical data (populations, business establishments, hybrid and OECD).

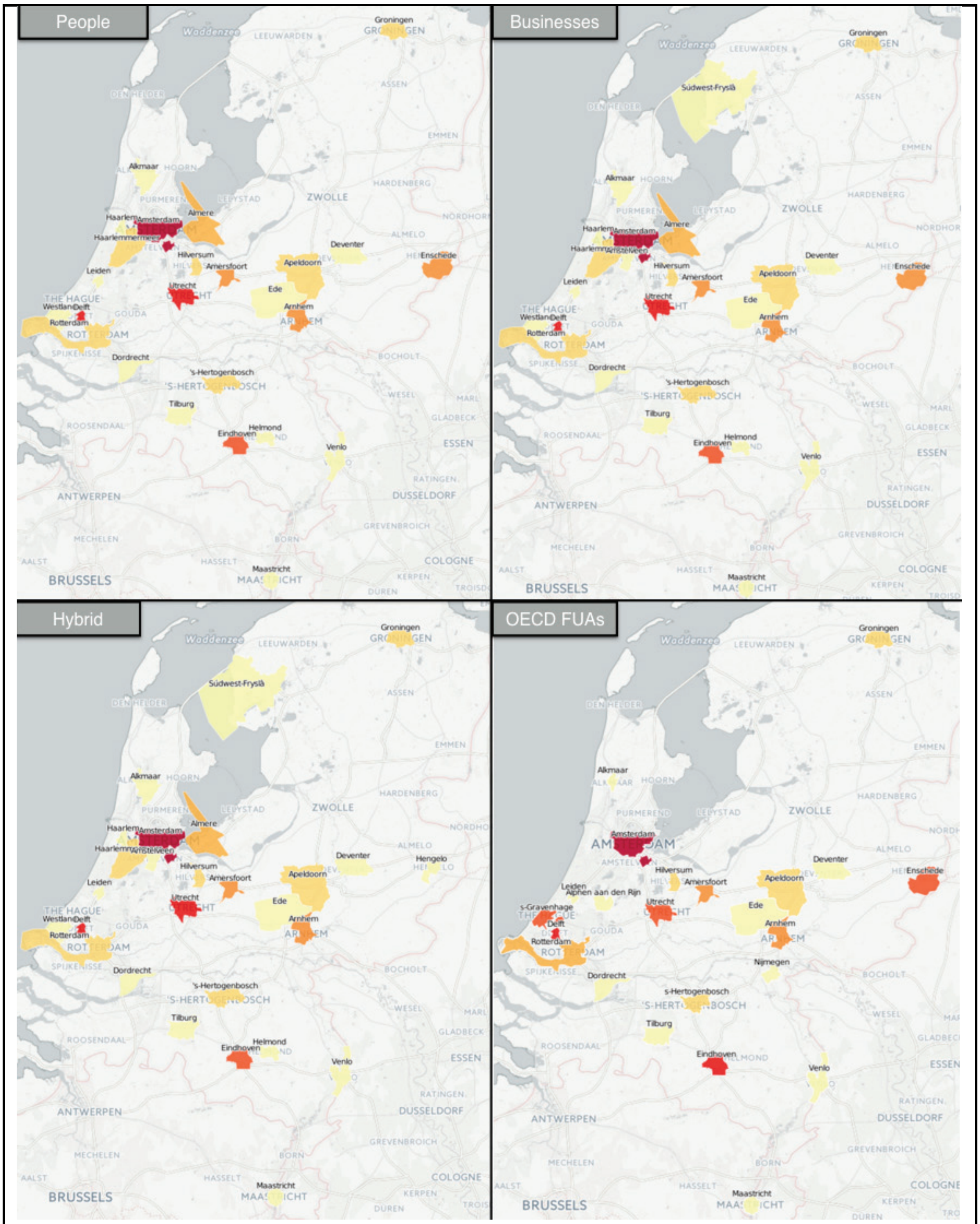


Figure 8. Amount of RVO project subsidies mapped to the dynamically delineated FUAs defined based on the CBS open statistical data and OpenStreetMap boundaries.

In the second step, geographical coordinates of projects are mapped to these Urban Areas. Fig. 8 shows the result of the mapping where frequency of the projects is highlighted: (again) the darker the color, the higher the number of awarded projects in that area.

As can be seen when comparing Figure 7 and Figure 8, by far not all (Functional) Urban Areas have projects. But more importantly, the different ways the Urban Areas are defined, lead to different outcomes. Using the OECD FUAs (bottom-right map in Figure 8), or the population density based FUA (top-left map), one would miss some of the relevant areas.

We plan to improve the underlying algorithm to be able to closely reproduce the OECD FUA. Moreover, we plan to use geographical data from GADM, instead of OpenStreetMap, which we used for this case study, to improve the mapping of projects (cf. see Figure 4). In the case study we identified hybrid FUAs by selecting the 35 areas that contain most population and businesses. This decision was made in order to compare to the 35 OECD FUA areas. A selection based on more criteria and more or less than 35 areas will yield outcomes that are more relevant for different research questions and possibly also an outcome that is closer to the OECD FUAs.

After identifying the spatial dimension of the project consortia, one may aim to find out whether the 'geographical containers' share socio-economic or other factors, in which they differ from less innovative areas. In this example, we only mapped the geo-location of the main applicant in each of the consortia. In a next step, we will also include the geo-location of the other partners in the projects.

Category services

As datasets may use different category systems for the attributes, linking data requires a mapping of these category systems or 'vocabularies' (Figure 9). A good example are the different category systems that are used for classifying research fields, e.g., in the Web of Science and in OECD R&D statistics; or different geographical schemas used for describing locations e.g. ISO country codes such as ISO Alpha-2, Alpha-3, and Numeric Country Codes. A *category service* would enable the data user to select which classification he/she wants to use. And the system would then do the mapping between the different classifications. For this, we deploy existing vocabularies available on the web. One can also think of other classification schemes that can be mapped, e.g., of professions, of jobs, of types of organizations, and so on. As many developments are taking place, the SMS platform may use what is available within the RISIS project work is done on classifications of companies, and of research organizations. The RISIS metadata system will be of help here.

Improving data quality

Linking can also be used to *improve quality* of the data. Linking the two sets may increase the number of variables, but also may reveal discrepancies in variable values, and the user should then be able to decide what the more reliable source is. Quality improvement follows from detecting value differences or similarities between datasets. *Quality assessment* using among other *provenance* will be implemented too: What was done with the data, and how. This should be transparent for the user.

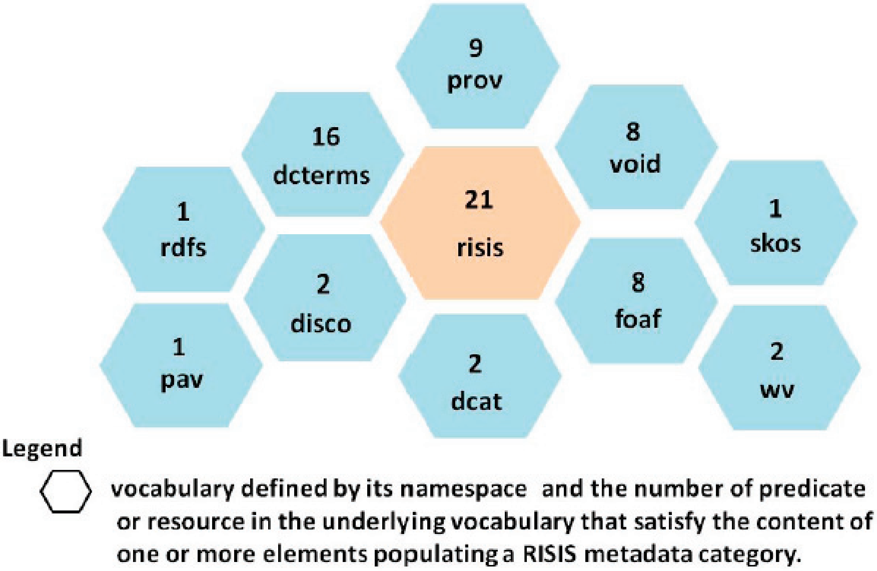
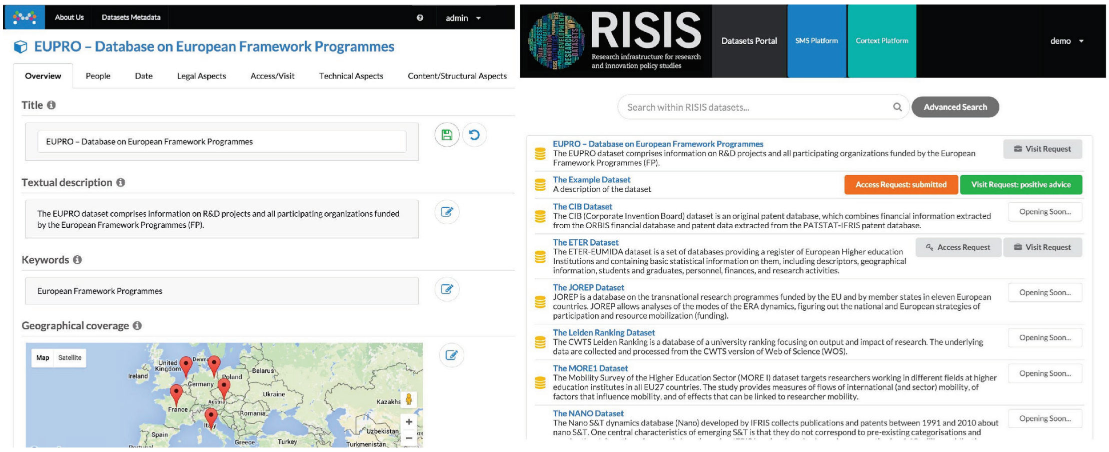

9
prov
16
dcterms
8
void
1
rdfs
21
risis
1
skos
2
disco
8
foaf
1
pav
2
dcat
2
wv
Legend
vocabulary defined by its namespace and the number of predicate or resource in the underlying vocabulary that satisfy the content of one or more elements populating a RISIS metadata category.

Figure 9: Integrating vocabularies

Metadata

The platform provides access to a variety of datasets, of which some are open, some are proprietary and require e.g., subscription, and some are confidential. The platform offers a metadata system, which is also linked to open data in order to have advanced search facilities even on data that are not publicly available.

The metadata system is also a tool to support data integration, due to the fact that the dataset owner is stimulated to use URLs in the metadata (Figure 10). And it is supported by the category services discussed above. (For more details: Idrissou et al (2016).



EUPRO – Database on European Framework Programmes

Title

EUPRO – Database on European Framework Programmes

Textual description

The EUPRO dataset comprises information on R&D projects and all participating organizations funded by the European Framework Programmes (FP).

Keywords

European Framework Programmes

Geographical coverage

EUPRO – Database on European Framework Programmes
The EUPRO dataset comprises information on R&D projects and all participating organizations funded by the European Framework Programmes (FP).

The Example Dataset
A description of the dataset

The CIB Dataset
The CIB (Corporate Invention Board) dataset is an original patent database, which combines financial information extracted from the ORBIS financial database and patent data extracted from the PATSTAT-IFRIS patent database.

The ETER Dataset
The ETER-EUMIDA dataset is a set of databases providing a register of European Higher education Institutions and containing basic statistical information on them, including descriptors, geographical information, students and graduates, personnel, finances, and research activities.

The JOREP Dataset
JOREP is a database on the transnational research programmes funded by the EU and by member states in eleven European countries. JOREP allows analyses of the modes of the ERA dynamics, figuring out the national and European strategies of participation and resource mobilization (funding).

The Leiden Ranking Dataset
The CWTS Leiden Ranking is a database of a university ranking focusing on output and impact of research. The underlying data are collected and processed from the CWTS version of Web of Science (WOS).

The MORE1 Dataset
The Mobility Survey of the Higher Education Sector (MORE I) dataset targets researchers working in different fields at higher education institutes in all EU27 countries. The study provides measures of flows of international (and sector) mobility, of factors that influence mobility, and of effects that can be linked to researcher mobility.

The NANO Dataset
The Nano S&T dynamics database (Nano) developed by IFRIS collects publications and patents between 1991 and 2010 about nano S&T. One central characteristics of emerging S&T is that they do not correspond to pre-existing categorisations and

Figure 10: The RISIS/SMS metadata system

The metadata creator/editor

The RISIS dataset holders describe their datasets in a detailed, consistent and uniform way, store the description and sometimes modify the stored metadata. On top of that, they have to model the description in RDF, a data model that they are often not familiar with. To stimulate non-Semantic Web users to generate valid RDF metadata descriptions, we designed a novel user-friendly editor. The editor covers the following categories of metadata:

- *General*. For an overview of a dataset, the metadata provides information that specify the source from which the dataset was derived, the title of the dataset, the language of the dataset, its version, its geographical coverage, its textual description, the keywords which describe the main topic of the dataset, the home page where further information about the dataset could be found or other related pages that are not the dataset's homepage but yet provide information on the dataset. A nice feature is the use case property, which provides access to published work derived from the RISIS datasets.
- *Temporal aspects*. Various temporal related information associated with an event in the life-cycle of a dataset is covered in the vocabulary. Such type of information concern data collection date, time coverage of the data itself, created date, issued date, modified date, started date and ended date.
- *Content*. To address the content of a dataset, metadata provides ways for users to have an idea on the type of resources present in the dataset by describing the entity types present in the dataset, by providing example of a resource and, by providing a sample of the dataset with all its properties. Likewise, it provides tables of all abbreviations, classifications and vocabularies used in the dataset and their respective descriptions.
- *Structure*. Information about the structure of a dataset describes the name of the tables in the dataset, the number of records contained in a table, the different attributes covered by a table. It also specifies whether the dataset is a partition. If it is a partition, it provides information on whether the partition is an entity type-based partition or a attribute-based partition.
- *People*. The metadata describes the contributor, the creator or the publisher of a data by giving their respective contact information and, names in different flavors: short, long and preferred.
- *Technical aspects*. RISIS metadata provides information about the dataset model used. This informs on whether the dataset follows the traditional tabular model (Relational, Spreadsheet, etc.) or the graph model (RDF). It also covers other information such as the format and the size of the dataset.
- *Access*. To inform the data consumer on how to access or query the data, the metadata provide information such as the opening status which notifies whether the data is open for visit, access type which specified whether the data can be visited, requested or both or whether the data is access free. In addition, it provides access URL which is information on the landing page, feed, SPARQL endpoint or other type of resource that grant access to the distribution of the dataset and, the data download address which is information on the location of the dataset for download.
- *Legal aspects*. The legal aspects of a dataset is covered by the RISIS metadata through a license which explicitly determines the terms under which a dataset can be used, rights which provides information such as property and intellectual rights associated with the data, terms of use which describe non-binding conditions, access

conditions and visit conditions which respectively describe the conditions in which end-users can access or visit a dataset, and Non-Disclosure Agreement (NDA) which specifies conditions of access to confidential information which would need signing a NDA with the dataset holder(s).

In order to provide a Linked Data-based User Interface (UI), we adopted our component-based UI framework called Linked Data Reactor (For more details: Khalili et al (2016). We implemented all the following features which were extracted from the requirement analysis and interviews with RISIS dataset holders:

- *Render metadata properties in different categories*. The user interface in designed in a modular way: It groups metadata information in different categories for separation of concerns.
- *Avoid presenting to the user technical metadata properties*. To emphasize on the generic aspect of the interface, we avoid displaying technical terms attached to certain properties. For example, instead of displaying technical metadata properties such as RDF Dump, Access URL (Uniform Resource Locator), and Byte Size, the interface respectively displays user-friendly alternative words such as Data Download Address, Access Address and Dataset Size.
- *Support metadata properties with hints*. To understand the meaning of the metadata properties provided by the interface, a hint is associated to each metadata property. The hint provides a description for the property to avoid confusion or ambiguity.

The workflow

From the users' perspective, the platform does two things. Firstly there is the workflow to identify data needed by the user to do a research project. This goes from identifying the entities and the variables (properties) needed. Through the metadata search the relevant datasets can be selected. If access can be given, steps follow like classification matching and disambiguation, and then the data can be integrated. The workflow is represented in Figure 11.

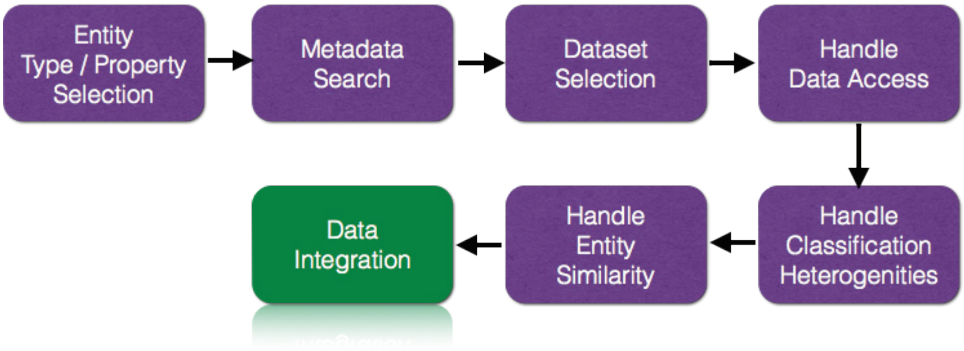


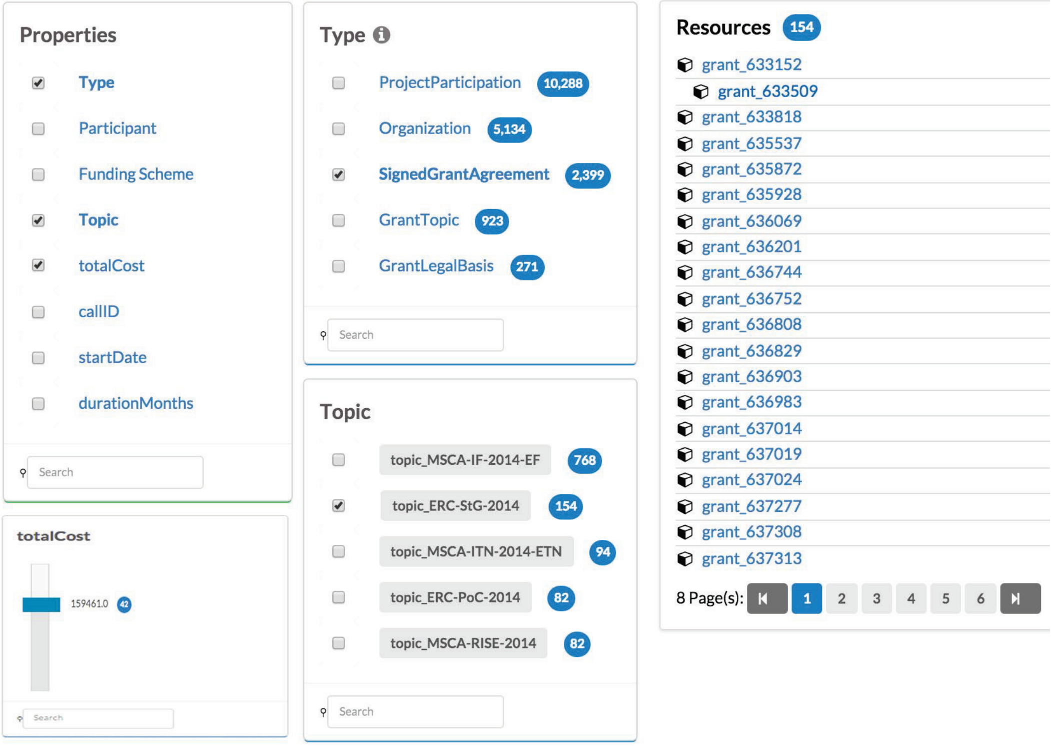
Figure 11: The users workflow

Secondly, when the integrated data are available, the user wants to have a specific dataset to do the analysis and visualization. (Standard) queries are provided to get the required data into the required format. This sounds simpler than it is, but experience with other data integration platforms show that the user needs support by data specialists to query the platform. This suggests that SMS / RISIS is indeed more an infrastructure than a tool. The output can have various formats, to enable deployment of general or specific analytical tools.

The faceted browser

The *faceted browser* enables exploration of datasets, in order to get a more qualitative feeling for the data. Figure 12 depicts faceted browsing of a project grant dataset. Users can explore the properties of resources in a dataset (top-left facet in Figure 12). Each property that is selected introduces a new facet in which users can select a specific value or range of possible property values, and thereby filter a subset of the resources (rightmost facet in Figure 12) based on specific properties. In Figure 12 the user has filtered resources (i.e., data element; individual grants) based on specific values for properties *totalCost* of the grant (bottom-left facet), grant *Type* (mid-top facet), and the *topic* of grants (mid-bottom facet). A subset of 154 out of 19,145 total resources in the dataset is filtered out and shown in the rightmost facet in Figure 12 based on the applied filters. Individual resources can be investigated in depth by clicking on their name, e.g., "*grant_633152*".

Faceted browsing is particularly useful when users want to explore the dataset a dataset via multiple entry points, or when users do not know what they are looking for beforehand. It allows users to explore the space of potential items by choosing the refinements, i.e., filtering of data properties, in any order. This helps users to analyze the data and prepare for a large-scale quantitative study of the dataset, or a study in which multiple datasets are linked together based on the properties previously explored in the faceted browser.



Properties

☑ Type
☐ Participant
☐ Funding Scheme
☑ Topic
☑ totalCost
☐ callID
☐ startDate
☐ durationMonths

totalCost

159461.0 42

Type

☐ ProjectParticipation 10,288
☐ Organization 5,134
☑ SignedGrantAgreement 2,399
☐ GrantTopic 923
☐ GrantLegalBasis 271

Topic

☐ topic_MSCA-IF-2014-EF 768
☑ topic_ERC-StG-2014 154
☐ topic_MSCA-ITN-2014-ETN 94
☐ topic_ERC-PoC-2014 82
☐ topic_MSCA-RISE-2014 82

Resources 154

grant_633152
grant_633509
grant_633818
grant_635537
grant_635872
grant_635928
grant_636069
grant_636201
grant_636744
grant_636752
grant_636808
grant_636829
grant_636903
grant_636983
grant_637014
grant_637019
grant_637024
grant_637277
grant_637308
grant_637313

8 Page(s): 1 2 3 4 5 6

Figure 12. An example of SMS faceted browser

Another use of the faceted browser is searching for information for more qualitative studies. The linked nature of the data enables users to search for rich information about the resources they are interested in. For example, each of the 154 resource elements (*grants agreements*) shown in the right facet of Figure 12 can be clicked on, and this shows the title of the grant, its participants, its duration, legal basis, funding scheme, objective, participants, topic, etc. The faced browser allows a researcher to filter the 154 grants in Figure 12 based on their total cost, and then find details that helps the researcher to perform a qualitative study of grant agreements in a specific cost-range in terms of legal status, topic of research, objectives, etc.

Analytical tools

The main output of the SMS platform will be a set of services to access to the enriched and integrated data for further analysis.

Output can be fed into several analytical tools, among others the RISIS-Cortext platform (www.Cortext.fr) for analysing textual data. RISIS-Cortext proposes several types of analysis, going from basic statistics to fine-grained network community analysis, made available to the user according to the chosen corpus type. A specific SMS interface is under development to connect the SMS platform to the Cortext platform. Output of the SMS platform is also available in formats that enable further analysis with the common tools for statistical analysis, network analysis and so on.

4. Conclusions

Technological advances have stimulated the development of natural science (De Solla Price 1984). New data integration and data enriching infrastructures may do the same for the social sciences (Christakis 2013).

We described the architecture and functions of the SMS platform, with emphasis on the use of open data, on data integration, and on geo-location. We expect that the SMS platform will enable research within the STI field that was not possible before. Existing STI studies are either *rich* but small scale (qualitative case studies) or large scale and *under-complex.* The SMS platform introduces support for the integration, analysis, and visualization of large datasets. This in turn allows for more large-scale STI studies involving the use of more interlinked and enriched data, and many more variables than traditionally has been the case.

Two case studies illustrate the practical use of the SMS platform's features to answer research questions. Many parts of the platform are already implemented and tested. Currently we are finalizing the beta-version of the platform, and according to the current planning, the SMS-platform will be available for users before the end of 2016.

5. References

Beek W, Rietveld L, Schlobach S, Van Harmelen F, LOD Laundromat; Why the semantic web needs centralization (even if we don't like it). IEEE Internet Computing, March-April 2016

Christakis N, Let's shake up the social sciences" *New York Times* (July 19, 2013),

Ciccarese P, S. Soiland-Reyes, K. Belhajjame, A. J. Gray, C. Goble, and T. Clark. Pav ontology: provenance, authoring and versioning. *Journal of biomedical semantics*, 4 (2013) 1,1-22,

Daraio C, M. Lenzerini, C. Leporelli, H. F. Moed, P. Naggar, A. Bonaccorsi, and A. Bartolucci. Data integration for research and innovation policy: an ontology-based data management approach. *Scientometrics*, 1-15, 2015.

de Solla Price D, The science/technology relationship, the craft of experimental science and policy for the improvement of high technology innovation. *Research Policy* **13** (1984) 3-20

Groth, P, A. Loizou, A. J. Gray, C. Goble, L. Harland, and S. Pettifer. Api-centric linked data integration: The open PHACTS discovery platform case study. *Web Semantics: Science, Services and Agents on the World Wide Web* **29** (2014) 12-18

Gurney T, Horlings E, van den Besselaar P, Author Disambiguation Using Multi-Aspect Similarity Indicators. *Scientometrics* **91** (2012) 435-449

Idrissou, AK, Khalili A, Hoekstra R, van den Besselaar P, Managing metadata relevant for research and innovation studies: The RISIS case. Paper *Whise workshop* 2016

Khalili A, A. Loizou, and F. van Harmelen. Adaptive linked data-driven web components: Building flexible and reusable semantic web interfaces. *Extended Semantic Web Conference* (ESWC) 2016.

Sandström, U., & Sandström, E., The field factor: Towards a metric for academic institutions. *Research Evaluation* **18** (2009) 243–250.

Van den Besselaar, P., L. Stout, X Gou, Predicting panel scores by linguistic analysis. In: *Peripheries, frontiers and beyond; Proceedings of the 21st International Conference on Science and Technology Indicators*. València (Spain), September 14-16, 2016.

Van den Besselaar & Sandström, What is the required level of data cleaning? A research evaluation case. *Journal of Scientometric Research* **5** (2016) 1, 7-12.

Van den Besselaar P, Schiffbanker H, Sandström U, Holzinger F, Polo L, *Explaining gender bias in grant selection – the ERC starting grants case*. Paper presented at the Conference on Gender and Higher Education, Paris, September 2016