The Semantically Mapping Science (SMS) Platform: Towards an Open Linked Data Infrastructure for Social Science Research

Ali Khalili [a], Al Idrissou [a,b], Klaas Andries de Graaf [a], Peter van den Besselaar [b,*] and Frank van Harmelen [a]

[a] *Department of Computer Science, Vrije Universiteit Amsterdam, the Netherlands*
E-mails: a.khalili@vu.nl, o.a.k.idrissou@vu.nl, ka.de.graaf@vu.nl, frank.van.harmelen@vu.nl
[b] *Department of Organization Sciences, Vrije Universiteit Amsterdam, the Netherlands*
E-mail: p.a.a.vanden.besselaar@vu.nl

Abstract. Social phenomena are generally complex. Understanding them, and designing public policies that may affect them, requires integrating and analyzing data from multiple sources. Currently, social research is mostly either rich but small scale (qualitative case studies) or large scale and under-complex (because it generally uses a single dataset - often a survey or administrative data). Progress in the social sciences depends on the ability to do large-scale studies with many variables specified by relevant theories: There is a need for studies which are at the same time big and rich, and this requires high quality linked and enriched data, that can be accessed through user-friendly interfaces. The Semantically Mapping Science (SMS) platform, presented in this paper, is a user-centric platform for data enrichment, integration, exploration and analysis with focus on open access to research data and services to tackle this challenge. We show the added value of the SMS platform through a number of illustrative use-cases. The SMS platform focuses on the data needs of researchers, policy makers, and managers in the area of science, technology, and innovation policies, but it generalises to data in other social science domains.

Keywords: Semantic Web, Linked Data, eScience, Social Science, Science and Innovation Studies, STI, Research Infrastructure

1. Introduction

In this paper, we present the Semantically Mapping Science (SMS) platform as an open linked data infrastructure for integrating and enriching heterogeneous public and private data, varying from tabular statistical data to unstructured data found on the Web, in an innovative and meaningful way. SMS is built as an open source platform[1] and is available online at http://sms.risis.eu.

The platform aims to support social research, and in its current implementation it focuses on the field of science, technology and innovation (STI) studies, an interdisciplinary field between the social sciences and humanities.[2] It covers fields from the economics of science and innovation up to the history and philosophy of science [1], and it depends on a large variety of heterogeneous data: *structured* and *unstructured*, *qualitative* and *quantitative*.

The SMS platform has the following distinctive characteristics: (i) it supports all steps in the process of data input, data linking, data enrichment, data selection, and data retrieval in a standard format that can be used in a variety of analysis and visualization tools; (ii) it hosts and integrates heterogeneous data that play a role in social science research; (iii) the metadata system and the user-friendly interface support diverse categories of users in finding, exploring, and download-

*Corresponding author. E-mail: p.a.a.vanden.besselaar@vu.nl.
[1] https://github.com/risis-eu/sms-platform

[2] Which means that most of the datasets included relate to science and innovation studies.

ing segments of the data needed for answering their research questions.

From the user's perspective, the SMS platform has the following functions: (i) It includes a large number of datasets with information on the central entities of interest in science and innovation studies: people, organizations, projects, outcomes (papers, patents), as well as statistical information related to these entities. (ii) The different datasets are (at the moment partly) linked, using an advanced tool for disambiguating names of persons and organizations. (iii) Through a powerful and flexible geo-location tool, the entities can be related to (geo)statistical data such as average income, average educational level, and so on. Multi-level geo-boundaries are included, and the researcher can select the level of granularity that fits the research question. (iv) An annotator tool enables to analyze textual data. (v) The faceted browser enables the user to browse through the data sets, visualize the data in various ways, select the data needed for his/her study and retrieve those for further analysis.

In this paper, we summarize the architecture and main features of the SMS platform, and present a number of use cases. The paper has the following structure:

- *Technical Architecture*. Section 2 gives an overview of the main technical components of the SMS platform (based on [2]).
- *Data Curation*. Section 3 describes the process of ingesting data into the platform and how metadata are created and edited for the imported datasets (based on [3]).
- *Semantic Enrichment of Data*. Section 4 presents two methods for enriching existing textual data on the platform (based on [4]).
- *Data Linking*. Section 5 describes the SMS approach for context-sensitive entity linking and disambiguation (based on [5, 6]).
- *Querying and Browsing of Data*. Section 6 presents the LD-R framework developed for providing flexible user interfaces for querying and browsing data(based on [7–11])
- *Use Cases*. Section 8 provides use cases that show how the SMS platform can be used by social science researchers to answer different types of research questions.

For each of these aspects we summarise per section the main lessons for Semantic Web technology that we extract from our experience in building and evaluating the SMS platform.

Most of the above functions have been described in earlier Semantic Web and Social Science conference publications. This article brings together these individual contributions, it provides an overarching description of the SMS platform, and it distills a number of lessons on the use of Semantic Web technology for supporting scientific data management. It is also the first to describe a number of use cases in some detail.

2. Technical Architecture

As shown in Figure 1, the SMS platform consists of three main layers: *data layer*, *service layer* and *application layer*.

The data (curation) layer deals with data acquisition, conversion, storage and access plans. The data available on the platform originate from open data on the Web, available public datasets (not published on the Web) as well as private (or restricted) datasets. The standard data representation used in the platform is *RDF*. In the process of converting data to RDF, for private datasets either only metadata is converted (in case of sensitive data e.g. personal data) or data is converted together with strict access rules (in case of subscription-based datasets such as the Web of Science dataset). The converted data is then stored in a triple store.

The Service layer provides a set of Web services on top of the triple store to allow developing applications. The SMS platform includes services for data curation, data linking, semantic enrichment of data as well as querying and browsing data.

The application layer is the terminal for end-users, allowing them to interact with the platform that offers user interfaces to facilitate a variety of Human-Data Interactions.

Lesson Learned. The three layer architecture from Fig. 1 is a repeat of the experiment with the OpenPhacts platform[3] for pharmacology data, reported in [12]: a *data layer* for ingestion, conversion and storage, a *service layer* for entity resolution, querying and inference, and an *application layer* for user-oriented functionality and interfaces. It confirms that this three layer architecture generalises across two areas of science with very different characteristics.

In the subsequent sections, we elaborate on these three layers of the SMS technical architecture.

[3] https://www.openphacts.org/

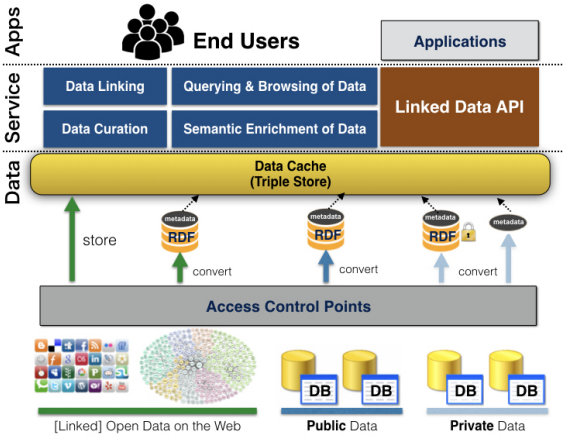

Fig. 1. The technical architecture of the SMS platform [2].

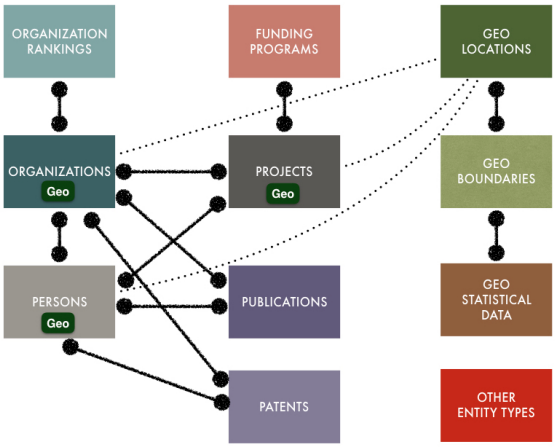

Fig. 2. The main entity types involved in the SMS system (solid lines represent the explicit links between the entities and dashed lines represent the implicit links between the entities derived after the geo-enrichment of data).

3. Data Curation

At its conceptual model, the SMS platform employs an entity-centric approach to interlink heterogeneous datasets in the STI domain. After analysis of a number of representative datasets, we have designed the following schema to interlink heterogeneous datasets in the STI domain (see Figure 2): 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 that needs to be supported by the SMS infrastructure.

Among the datasets included in the SMS data store are several datasets on research intensive organizations: GRID[4], ETER (the European Tertiary Education Register)[5] and its recent extension OR-GREG[6], and the Leiden Ranking[7]. Output datasets are included such as USPTO[8], (parts of) bibliometric datasets, and recently developed open access datasets as OpenAIRE[9]. Datasets with research projects are included too, currently mainly at the European level (CORDIS[10]) but extension with project data from national research funders is foreseen. To support geolocation, several datasets are available such as the OECD FUA[11] data, GADM[12] administrative boundaries, OpenStreetMap[13], and Flickr[14]. Finally, data from statistical offices such as CBS[15] (Statistics Netherlands) are included in the data store.

To ingest the above datasets into the platform, we need to deal with issues such as heterogeneity of data formats and structures, the access level of data, provenance information as well as semantics of data. In the following subsections we describe how we address these data curation issues.

3.1. Semantically Neutral RDF Conversion

In order to provide a *semantically neutral RDF* conversion of data (by which we mean a conversion that preserves the original semantics of data), we follow a *simple syntactical RDF conversion* that allows to add user specific context to the data at usage time. This assumes no commitment to any vocabulary, maintains as faithfully as possible the CSV columns as RDF properties, and maintains the provided entity type. Current conversion methods prematurely commit to a context at conversion time, while the original data could as well fit many more contexts that could not be prede-

[4] https://grid.ac
[5] https://www.eter-project.com
[6] https://risis-eter.orgreg.joanneum.at
[7] http://www.leidenranking.com
[8] https://www.uspto.gov
[9] https://www.openaire.eu
[10] https://data.europa.eu/euodp/data/dataset/cordisH2020projects
[11] http://www.oecd.org/cfe/regional-policy/functionalurbanareasbycountry.htm
[12] https://gadm.org
[13] https://www.openstreetmap.org
[14] http://code.flickr.net/2011/01/08/flickr-shapefiles-public-dataset-2-0
[15] https://www.cbs.nl/en-gb/our-services/open-data

fined or anticipated at conversion time. In the conversion stage, these commitments add unnecessary complexity - constraints to specific interpretation - to RDF graphs that might not need it. We argue that, enforcing the data with ontological commitments increases the probability of information loss, and it adds terminology, semantic and context bias to the converted data. For these reasons, we advocate conversion neutrality instead as it maintains data integrity and keeps it bias free. Conversion neutrality dictates keeping the implicit expressivity of the original non-RDF source by postponing the addition of semantic complexity: *not* link them to other ontologies, and only do this at the time when using the data for a particular research question.

This semantically neutral idea is supported on the platform as Access Control Points (ACPs) where data providers connect their data servers directly to the SMS platform so that data is fetched on demand and neutrally converted as documented at http://acp.api.risis.eu. The documentation highlights that each ACP is required to implement four REST APIs: METADATA, ENTITY-TYPES, ENTITY and ENTITIES. The first one returns relevant metadata for understanding the intended meaning of the data source. The second API lists all entity-types documented by the data source. The third API allows for the request of a specific entity using the entity's identifier and type. Using only the entity type, the fourth and last API allows for the request of entities of a specific types. These aforementioned APIs on the SMS platform, facilitate the automation of data conversion into semantically neutral RDF data. They also provide a mechanism to coordinate access to data based on *user role* and the *dataset owner's requirements* by specifying data access at the levels of dataset, entity and properties.

Lesson Learned. Adopting a semantically neutral RDF conversion approach in which data is converted into a neutral schema, allows building flexible applications capable of working with multiple semantics that are tailed to specific user groups and use-cases.

3.2. Dataset Description and Discoverability

Within the SMS platform, metadata helps potential users of a dataset to decide whether the dataset is appropriate for their purposes or not. It has been shown that research publications that provide access to their base data yield consistently higher citation rates than those that do not [13]. The SMS platform hosts a collection of various heterogeneous datasets that are not all publicly accessible due to privacy issues. To access these datasets, one needs to be granted an access request. This administrative detour that a researcher has to endure prior to detecting which dataset to use for a particular research question can reduce the number of SMS dataset visitors. Therefore, to mitigate the problem of access to private datasets and to attract more users to visit and use (and cite) the SMS datasets, metadata features are considered as a crucial component.

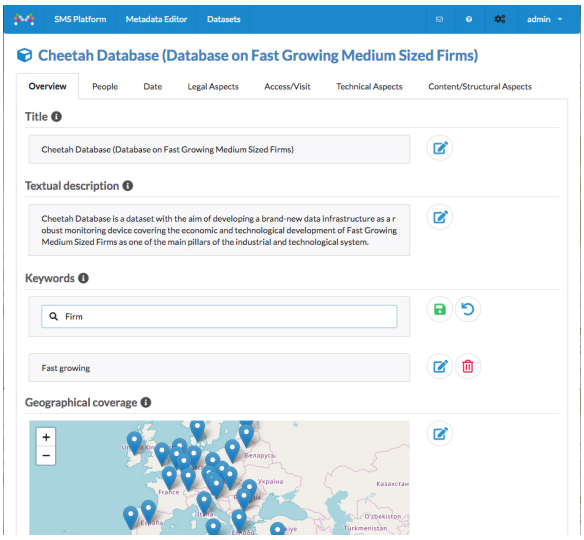

Fig. 3. An screenshot of the SMS metadata editor.

Although metadata can be used for tasks such as data validation or data conversion, the SMS metadata is particularly meant to meet five requirements [3]: 1) facilitate dataset descriptions *displayed* at a user interface level, 2) provide information *guiding the use of data*. The detailed information about the datasets is required in order to 3) help users to get an *in-depth understanding* of the data at hand, in such a way that they could easily identify how the data should be interpreted, used, or linked to other data – without having to inspect the data itself. Facilitate trust by 4) providing *details about the quality of the underlying data*. Finally, a requirement is to 5) facilitate *simple and advanced search for relevant data*. This is considered to be a crucial task for data discovery and link discovery across datasets. To illustrate, a simple search query could be to return all datasets that are about higher educations. In a complex search, a user could ask to retrieve all datasets that have information about higher education institutions in European capitals. To support

complex queries, the system exploits external background knowledge (e.g. from Geonames or DBpedia). In this way, the system makes use of additional information or knowledge that is not explicitly contained in SMS datasets. We divided the five requirements into the following metadata types: dataset overview, dataset temporal aspects, dataset content, dataset structure, person details, dataset technical aspects, dataset legal aspects, access and visit and data quality / used methodologies.

To develop the SMS metadata vocabulary, we investigated public dataset platforms and shared vocabularies (such as DataHub[16] and LOV[17]) that satisfy the requirements analyzed above. Since the SMS metadata cover a broad range of aspects (i.e. different types of information), it makes it difficult to find a vocabulary that covers all those aspects. Therefore, we decided to select a set of publicly shared dataset metadata vocabularies that were originally designed for one or more specific requirements which are included in the SMS platform. Figure 4 shows the metadata vocabulary designed to collect the metadata required on the SMS platform.

On top of the developed metadata vocabulary, the SMS platform provides a metadata service and application targeted at both data owners and data consumers. The metadata service API allows researchers to search for data, and have an understanding of the data without the need to directly access it. The metadata service is powered by an intuitive UI (cf. Figure 3) which allows dataset holders to describe their datasets in a detailed, consistent and uniform way, store the description and if needed modify the stored metadata.[18] The curated metadata are then accessible on RISIS dataset's portal available at http://datasets.risis.eu.

Lesson Learned. Designing and deploying a single metadata ontology that is uniform across datasets facilitates both exploration and integration of data. Furthermore, it facilitates differential access to data based on the user role and data owner's requirements.

4. Semantic Enrichment of Data

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

4.1. Named Entity Recognition

Named-entity recognition (NER) (also known as entity identification and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. Given a RDF dataset which has one or more attributes with textual values, SMS NER service extracts named entities from the text and more importantly connects the extracted entities to a knowledge graph or taxonomy (which can then provide more data about those entities). By default, SMS employs DBpedia Spotlight[19] service for NER. However, any arbitrary NER service can be plugged into SMS NER service as long as the output of the service is compatible with the SMS named entities annotation model. The annotator UI allows users to select a dataset, select a specific resource type and then a property that contains textual values. The annotations can either be stored in the original dataset or a separate dataset that reuses the identifiers from the original dataset. The UI also provides some advanced options for users to select custom NER APIs, select different languages and configure the annotator in terms of confidence level and stop words. Furthermore, to enable scalable annotations, users can select a batch size to increase the number of parallel requests sent to the NER API. As shown in Figure 6, SMS provides an interactive UI to give real-time feedback while a dataset gets annotated using the NER service[20].

4.2. 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 socioeconomic data. In order to exploit these linking and enriching possibilities, the SMS platform provides a variety of geo-services. The geo-services are based on a series of open geo-resources, such as GADM, OpenStreetMap and Flickr

[16] https://datahub.io/

[17] http://lov.okfn.org/

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

[19] http://www.dbpedia-spotlight.org/

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

People
dcterms:creator
dcterms:publisher
dcterms:contributor
skos:prefLabel
rdfs:label
foaf:name
foaf:familyName
foaf:givenName
risis:shortName
risis:fullName
foaf:mbox

DS Overview
pav:version
foaf:homePage
foaf:page
dcterms:spatial
dcterms:source
dcterms:title
dcterms:description
dcterms:subject
dcterms:language
risis:useCase

DS Structure Aspect
risis:table
risis:tables
risis:records
risis:attribute
risis:attributes
void:subset
void:classPartition
void:propertyPartition

DS Access-Visit
dcat:accessURL
void:dataDump
risis:accessType
risis:openingStatus

DS Content
void:exampleResource
void:vocabulary
void:class
risis:classes
risis:classification
risis:abbreviations
risis:datasetSample

DS Technical Aspects
risis:datasetModel
dcat:byteSize
dcterms:format

DS Methodology
dcterms:title
dcterms:description
risis:dQMethodology

DS Temporal Aspect
risis:dataCollectionDate
dcterms:temporal
dcterms:created
dcterms:issued
dcterms:modified
disco:startDate
disco: EndDate

DS Legal Aspects
dcterms:license
dcterms:rights
wv:norms
wv:waiver
risis:accessConditions
risis:visitConditions
risis:nonDisclosureAgreement

Fig. 4. The SMS metadata ontology with a view over vocabularies reused.

Fig. 5. Different levels of geo-boundaries used in the SMS platform.

geotagged data. By integrating these geo-resources, the service can give for an entity's address the geo-location up to 11 different levels of (administrative) boundaries. Figure 5 shows an example of these geo-boundaries for a given address in the Netherlands. One practical application we built for batch processing of addresses is a Google spreadsheet add-on[21] which chains Google Geocoding API with our geo-boundary services.[22] Given addresses in a spreadsheet are enriched with different levels of administrative boundaries. The users are then able to export the extracted geo-boundaries and process them in geodata analysis tools such as CartoDB.[23] We have also developed a user interface for automatic geo-enrichment of RDF datasets in the SMS platform. Similar to the NER UI, the interface allows users to select an existing dataset and geocode the whole dataset by selecting the right attributes in the dataset[24].

Lesson Learned. It is fully possible to use a variety of external services and knowledge bases for semantic enrichment of scientific data. The adoption costs of such external services and knowledge bases are nowadays sufficiently low, and their quality and reliability sufficiently high.

5. Data Linking

Entity resolution is an essential step in the use of multiple datasets on the semantic web. For more than a decade the Ontology Alignment Evaluation Initiative

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

[22] http://api.sms.risis.eu/#/Geo-Services

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

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

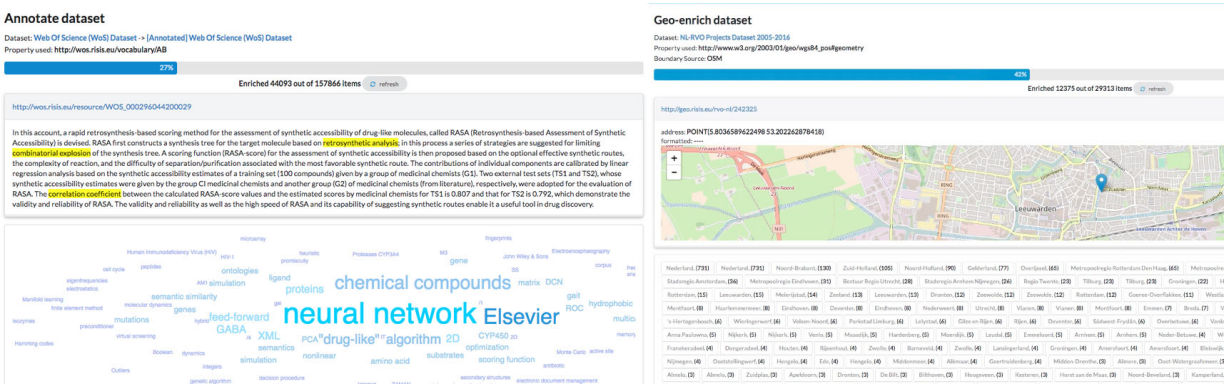

Fig. 6. Interactive UIs for semantic enrichment of data.

is dedicated in motivating researchers to finding better ways to address the problem. Nevertheless, since entity resolution algorithms are far from being perfect, the links they discover must often be human validated, which is both a costly and an error-prone process. Therefore, it is desirable to have computer support that can accurately estimate the quality of links between entities.

5.1. Link discovery

Linking between entities in different datasets is a crucial element of the SMS platform especially because it covers numerous and large datasets involving entities of interest to the field of technology and innovation studies. The SMS platform is innovative by combining three content-based standard link discovery approaches (string-based, number-based and geo-based) through the Lenticular Lens tools, which implements the *Lenticular Lens* [5] approach to generate, select and combine context-specific alignments, thereby allowing for future reuse.

Link discovery algorithms traditionally rely on the use of the `owl:sameAs` property to reflect the identity relationship. However, the semantics of this property entails that all properties and respective values are shared among matched resources without any further conditions. This is a problem because the identity relations in the complex world are context dependent [14], while the `owl:sameAs` makes it impossible to define contexts in which particular identity links hold. This often causes inconsistency through its cascading property and property-value sharing. For this, we argue in [5] that a possible approach is to tackle the problem at two fronts. We propose first, to *explicitly define the conditions (context) in which the identity holds* and second, to *explicitly list the properties (view) that are subject to property and value sharing,* hence making the remaining properties not subject to inference. For this, we distinguish between 3 types of properties. (1) PoI, the set of *Properties of Interest*, stands as an explicit declaration of properties for which `owl:sameAs` inferencing should apply within an alignment context. (2) Within the PoI, the subset of aligned-properties used for link discovery defines the conditions in which identities within the generated alignment holds. These properties are labelled as *Identity supporting Properties*. (3) The last set defines properties for which no inferencing should be applied (PnI) as opposed to the owl:sameAs semantics where all properties are subject to property-value sharing. In other words, it is the set of properties for which `owl:sameAs` semantics should not apply. This solution supports the idea that whether or not two entities should be considered equal depends not only on their intrinsic properties, but also on the purpose or task for which the entities are used.

Under the above proposed solution, directly using the `owl:sameAs` property is no longer feasible. Instead, SMS implements a *hierarchical singleton property model* that allows to define and use context dependent `your:sameAs` sub-property of `owl:sameAs`. As a result, a context specific alignment is now represented using a decorated alignment graph followed by annotated links. Together, these annotations account for how, why, when and by whom the links are generated, and help to select, combine and validate alignments in a context-specific manner. As shown in Figure 7, SMS exposes this approach through a UI[25] to

[25] Watch a screen-cast of the linking UI at https://youtu.be/CcffBlCBF54?list=PLo4YbUaRFSnwJ9XJvp6rlIMsaw_rfKT9C.

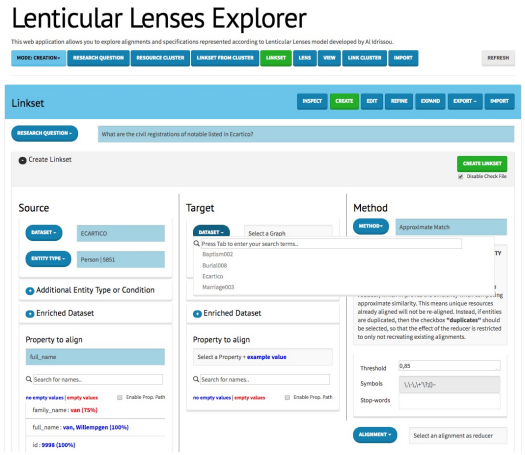

Fig. 7. Screen-shot of the SMS data linking UI.

allow end-users to create their own context specific alignments.

As an example, to study the performance of organizations, a researcher needs to align research organizations across datasets such as GRID[26] and OrgRef[27] that describe organisations across various countries including public and private R&D intensive organisations. For example, the 3M corporation, a large multinational organisation with a substantial patent portfolio, occurs in both datasets. GRID distinguishes between national 3M branches across six countries *3M (Canada), 3M (France), 3M (Germany), 3M (Israel), 3M(United Kingdom)* and *3M(United States)*, while OrgRef only refers to a single *3M* entity. Should these entities be designated as "the same" across these datasets? It depends. For a study that aims to compare multinationals at a global level, all branches of '3M' should be considered the same. Whereas, for a study that compares industrial structure across countries, the national branches of '3M' should be considered separately.

5.2. Link quality estimation

Matching entities within or between datasets is a crucial step for datasets integration in the semantic web, and thereby, the main approach available on the platform for data integration. An important amount of matching tool exists due to the rich literature in the semantic web on different ways to address the entity resolution problem. Depending on the size or the number of datasets involved, the use of these tools or the implementation of the approaches proposed in the literature lead to the generation of a frightening number of links awaiting for validation. Compared to link discovery, much less work has been done on link validation in terms of *"how to assess the quality of entity links once discovered?"*.

Non-automated evaluation methods for link quality are typically limited to either comparison with a *ground truth dataset*[28] (which is often not available), *manual work* (which is cumbersome and prone to error), or *crowd sourcing* (which is not always feasible, especially if expert knowledge is required). On automated link evaluation, two approaches exist. Both of them employ network metric to answer two research questions that end up composing two sides of the same coin. One looks at identity network structure and the symmetry of `owl:sameAs` for detecting erroneous links [15]. The other one, the one offered by the SMS platform through the Lenticular Lens Tool, also looks at the structure of an identity network for predicting the quality of the network as a whole [6]. The first approach has been tested on the Linked Open Data cloud where the network structure is combined with `owl:sameAs` symmetry representing an agreement between *both* data owners on the identity of the resolved resources to detect the degree of error of a discovered link. On the SMS platform we instead measure agreements between *multiple* data sources for estimating the quality of an identity network by investigating how a network bridge, density and diameter convey information on the quality of the identity network. This has been successfully tested on a number of datasets in the semantic web and in social science, and is currently being tested in the humanities domain. Both approaches indeed show the importance of the identity network structures for automatically assessing the quality of identity networks.

Lesson Learned. (a) The current Semantic Web methods for entity linking are too crude, since they rely on the owl:sameAs semantics, which is not context sensitive, and (b) the quality of current entity linking methods is not good enough for a high-precision domain such as e-science, and therefore efficient methods for human validation of the candidate links are needed. To address problem (a) we have developed the

The code for running the Lenticular Lens Tool is available at https://github.com/alkoudouss/alignments

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

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

[28] The comparison can be automated, but generating the ground truth is manual work unless it is synthesized

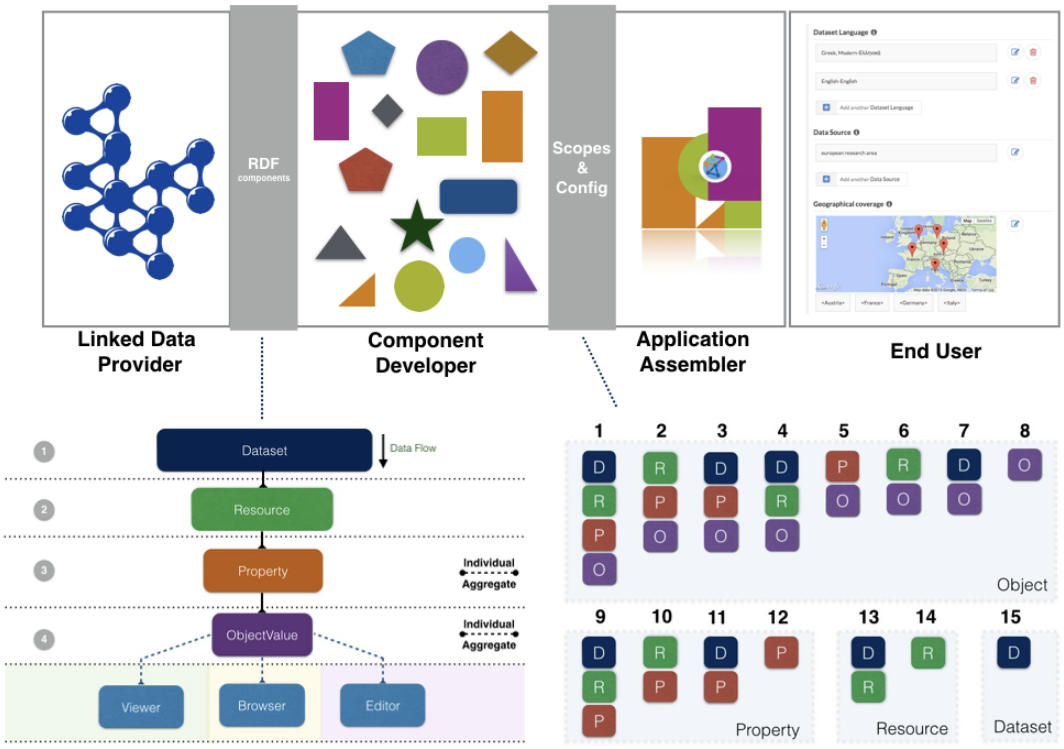

Fig. 8. Main elements of the adaptive LD-R Web Component's architecture.

notion of "lenticular lenses", which allow task-specific definitions of equality and identity (reported in detail in [5]); to address problem (b) we have developed efficient methods for automatic link quality estimation which can drastically reduce the human effort required for this task (reported in detail in [6]).

6. Querying and Browsing of Data

One of the objectives in developing the SMS platform was to enable non-Linked Data experts to smoothly query and browse RDF datasets without having the knowledge of SPARQL query language. In order to enhance the user experience while interacting with Linked Data on the SMS platform, we developed an open source software stack called Linked Data Reactor (LD-R) [7] that is available at http://ld-r.org. LD-R hides the complexity of interacting with Linked Data by employing a component-based software architecture. As shown in Figure 8, the LD-R components lifecycle addresses four primary types of stakeholders:

– *Linked Data Provider*. Since the LD-R approach focuses mainly on Linked Data applications, the provision of RDF-compliant data is an essential phase in developing the LD-R components. There are different stages [16] in Linked Data provision, including data extraction, storage, interlinking, enrichment, quality analysis and repair which should be taken into account by data scientists and Linked Data experts. Once the data and schemata are provided to the LD-R component system, the system allows Linked Data providers to better understand and curate the data when needed. For example, in the case of geo-coordinates, a map component can enable data providers to easily curate the outlier data within a certain geo boundary in a visual manner.
– *Component Developer*. Component developers are UX designers and Web programmers who are involved in component fabrication. There are two types of Web components developed in this step: a) *Core components* (a.k.a. Reactors) which abstract the underlying RDF data model. These components are built-in to the system, however can still be overwritten by developers who have proficiency in Semantic Web and Linked Data. b) *Community-driven components* which are higher

level components that exploit the core components. These components are either created from scratch or by remixing and re-purposing existing Web components found on the Web.

– *Application Assembler*. The main task of application assemblers is to identify the right components and configurations for the application; and to combine them in a way which fits the application requirements. Within the LD-R component system, the metadata provided by each Web component facilitates the discovery of relevant components. Sharing vocabularies on Linked Open Data allows assemblers to not only reuse components but also reuse the existing configurations and scopes published on the Web. For example, if there is already a suitable configuration which uses `foaf:Person` as resource type and `dcterms:description` as property URI, the assembler can reuse that configuration within his application.
  LD-R exploits a hierarchical permutation of the Dataset, Resource, Property, and Value (DRPV) components as *scopes* to select specific parts of the UI to be customized or personalized [9]. Each scope conveys a certain level of specificity on a given context ranging from 1 (most specific: DRPV) to 15 (least specific: D (Dataset). Scopes are defined by using either the URIs of named graphs, resources, and properties, or by identifying the resource types and data types. A configuration is defined as a setting which affects the way the UI components are interpreted and rendered (e.g., render a specific component for a specific RDF property or a specific RDF resource within a specific RDF graph). UI generation is handled by traversing the configurations for scopes, populating the configurations and overwriting the configurations when a more specific applicable scope is found.
– *End-User*. End-users experience interacting with the components to pursue goals in an application domain e.g. Science, Technology and Innovation studies. As such, they may request the development of new components or configurations in order to fulfill their requirements and are expected to provide feedback on existing components.

In the following section, we describe the LD-R faceted browser as one of the main features of LD-R incorporated in the SMS platform that enables end-users to visually query and browse Linked Data scattered over multiple knowledge graphs.

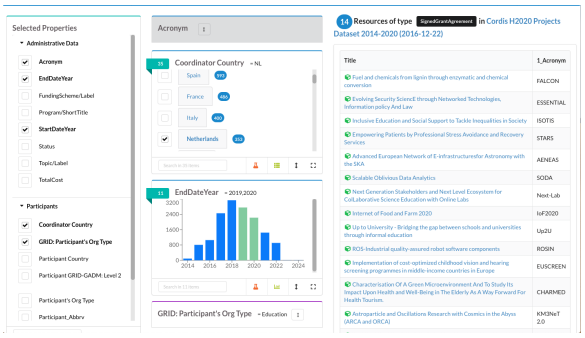

Fig. 9. An screenshot of the FERASAT faceted browser.

6.1. *FERASAT (FacEted bRowser And Serendipity cATalyzer)*

Serendipity, defined as the art of making an unsought finding [17] plays an important role in the emerging field of data science by increasing the chance encounter and thereby promoting unexpected knowledge discovery. The World Wide Web has provided a global information space comprising billions of connected documents. Early on, coming to the Web was a voyage of discovery, aimed at tracking how things connected, rather than just what they were about. However, most of the existing centralized nearest neighbor search approaches on the Web, such as Google, although very useful in finding explicitly relevant results, are killing serendipity by excessively limiting the encountering of unexpected information [18, 19]. On the other hand, the ever-growing amount of Linked Data publicly accessible and distributed on the Web increases the likelihood that some of the data, which will make an impact in our professional or private lives will come to us by chance—without searching it initially.

'Unsought discoveries' most often take place in the context of browsing unbounded data spaces; people immerse themselves in the items that interest them, meandering from topic to topic, and so on and so forth (i.e., the *Follow-Your-Nose* method [20] to traverse the given semantic links from a resource) while concurrently remarking interesting and informative information en route [21]. Therefore, flexible and intuitive browsing user interfaces (UIs) which support serendipity triggers, can increase the likelihood of accidental knowledge discovery on Linked Open Data (LOD).

SMS employs a novel faceted browsing UI called FERASAT [11] that provides an adaptive multigraph-based browsing interface to catalyze serendipity while browsing Linked Data. FERASAT aims to allow social science researchers who are not familiar with Linked

Table 1
List of the proposed serendipity-fostering design features [8].

Design Features Related to Observations	
F_1	Make surprising observations more noticeable.
F_2	Make errors in data more visible in order to detect successful errors easier.
F_3	Allow inversion and contrast.
F_4	Support randomization and disturbance.
F_5	Allow monitoring of side-effects when interacting with data.
F_6	Support detection and investigation of by-products.
F_7	Support background knowledge and user contextualization.
F_8	Support both convergent and divergent information behavior.
Design Features Related to Explanation of the Observations	
F_9	Facilitate the explanation of surprising observations.
F_{10}	Allow sharing of surprising observations among multiple users.
F_{11}	Enable reasoning by analogy.
F_{12}	Support extending the memory of user by invoking provocative reminders and relevance feedback.

Data technologies to explore a wide range of heterogeneous interlinked datasets in order to observe and interpret surprising facts from the data relevant to science, technology, policy and innovation studies. Figure 9 shows an screenshot of the FERASAT UI.

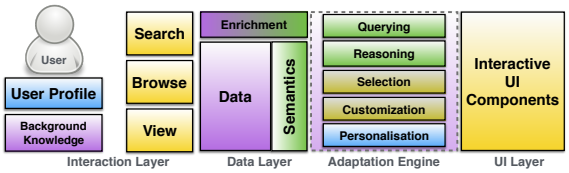

Fig. 10. Architecture of the FERASAT adaptive faceted browser.

Figure 10 depicts the architecture of FERASAT where related elements are color coded. The system provides three main modes of interaction with data namely search, browse and view. During the user interactions, based on the semantics of data and the given user context, the system adapts its behavior by rendering appropriate interactive UI components. In order to facilitate serendipitous knowledge discovery, FERASAT supports a set of serendipity-fostering design features (as listed in Table 1) to address the two main steps of serendipity namely a surprising observation followed by a correct interpretation. The details of each design feature together with its implementation are discussed in [8, 11]. To give you an example of these design features, F_3 suggests sharing of surprising observations among multiple users to increase the chance of interpretation of those observations: "A surprising observation done by user A, when correctly explained by user B, can result in positive serendipity."

In order to support this feature, FERASAT introduces the WYSIWYQ concept:

6.1.1. WYSIWYQ (What You See Is What You Query)

Existing Linked Data browsing user interfaces (UIs) allow users who are not familiar with Semantic Web technologies to explore interlinked datasets by generating SPARQL queries behind the scenes. This is analogous to the well-known WYSIWYG (What You See Is What You Get) paradigm, which generates the required markup in the background based on the user interactions. Nonetheless, and contrarily to the WYSIWYG approach where users are enabled to switch between the source code and the UI, there is no alternative in Linked Data browsing UIs to regenerate the browsing UI based on a given SPARQL query. To tackle this issue, FERASAT proposes a WYSIWYQ (What You See Is What You Query) query mode [10], a novel approach that allows people to interactively visualize a given SPARQL query as a browsing UI. The queries can then be further enriched or re-purposed by users in a faceted browser. As shown in Figure 11, the WYSIWYQ mode provides a two-way binding between SPARQL queries and the underlying faceted browsing environment by implementing the following mechanisms: a) componentize and customize a faceted browsing environment; b) identify, share and enrich SPARQL queries; c) validate SPARQL queries against a certain pattern; d) decompose a SPARQL query into a set of sub-queries which match a certain pattern; e) map SPARQL queries and their corresponding metadata to a set of customizable UI components.

Lesson Learned. A component-driven user interface model made based on RDF data model brings sub-

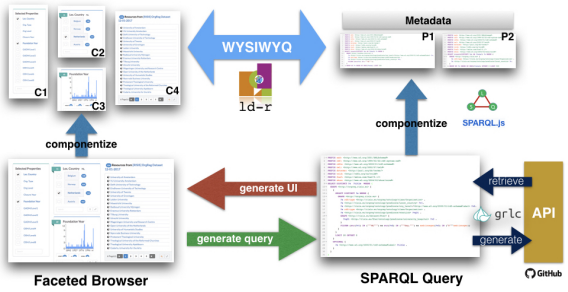

Fig. 11. The WYSIWYQ architecture.

stantial advantages for flexible viewing, browsing and editing of linked data. As was discussed in Section 6, reusing both components and configurations when designing a Semantic Web applications allows the easy introduction of novel features to adapt the user interface based on the context of the user and the underlying data semantics. Furthermore, it enables to translate SPARQL queries into a set of relevant Web components configured to interactively query the data. Finally, and alternative query-and-browse paradigm supporting serendipitous exploration increase the likelihood of unsought knowledge discovery in Linked Open Data.

7. Implementation

The SMS platform is implemented as an open-source custom fork of the LD-R framework that is based on ReactJS UI components and the FluxibleJS data flow framework for the front end, the Semantic-UI framework for themes as well as NodeJS for the back end. The source code is available on Github [29] together with the instructions how to install and use the platform. More information and video tutorials are available at http://sms.risis.eu website.

8. Use Cases

The SMS platform had 470 registered users (2018) varying from senior and experienced to junior researchers (professors, postdocs, PhD students), but also policy makers, librarians, project managers, scientific officers, etc.). In order to support use, several exemplary use-cases have been provided to show what can be done with the platform. These cover the topics such as investigating network structure of research organizations, analyzing change in the higher education system, analyzing the geography of innovation, evaluating research project portfolio's and explaining university rankings using characteristics of universities and the universities' environment.

A complete list of use cases is available at http://sms.risis.eu/usecases. In this section, we provide a brief summary of some of these use cases in the STI domain. All cases use the geo-enrichment function, as well as the querying and browsing functions. Cases 1, 2 and 3 use the data linking function. Data enrichment through the entity recognition function is central in use case 4.

8.1. *Adaptive Delineation of Functional Urban Areas using Linked Open Geo Boundaries*

Urban systems - a concentration of people, companies, research organizations and other activities - are crucial for future economic prosperity and quality of life for billions of people [22]. In order to investigate how these urban systems function, one needs to define the geographic boundaries. Administrative boundaries are often used, but administrative boundaries do not necessarily coincide with the urban system boundaries. For example, two cities very close to each other may function as one urban system. To solve this problem, the Organization for Economic Co-operation and Development (OECD), the European Commission (EC) and Eurostat have developed a new approach to classify what they call *Functional Urban Areas* (FUAs), based on population density (urban form) and work based commuting patterns (territorial organization). It is meant as a better definition of what are the urban areas in 28 OECD countries.

Even though this new approach may provide a better basis for an agreed definition of urban areas (UAs), it is at the same time rather rigid, as based on a fixed set of characteristics. To support researchers and policies, a more flexible approach may be needed where one can include additional factors not present in the OECD methodology, and where on can weight the factors differently: apart from population density and commuting patterns, other variables may be relevant to define the boundaries of geographical systems e.g., density of companies, or of transport infrastructures.

Therefore, SMS provides an adaptive approach for dynamic and multi-faceted delineation of (urban) areas, based on factors that are considered important from a policy or a theoretical perspective. This

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

adaptive definition of UAs demands integration of data from multiple up-to-date linked data sources. As shown in Figure 12, we deployed the SMS geo-enrichment services and applications to build a more flexible system for classifying geographic areas. The system combines openly available spatial and non-spatial resources on the Web.

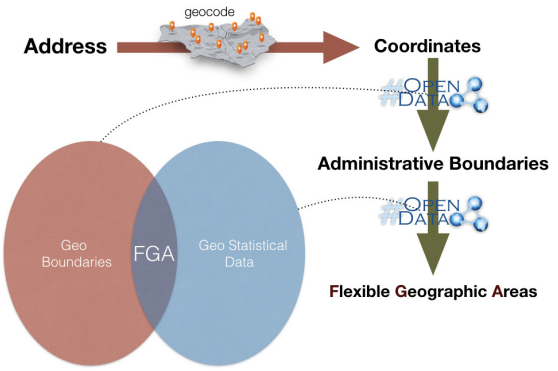

Fig. 12. The methodology to define flexible geographic areas.

In this use case [4] we show how using flexible geographic boundaries helps to answer the question (i) in which Dutch areas innovative activities are concentrated, and (ii) what the socio-economic and structural properties of these areas are. Recent research and innovation policies in the Netherlands focus on the so-called 'top sectors' of the economy, which were selected in a consultation of policy makers, representatives of the research system and entrepreneurs in the country. As a large part of total research and innovation funding is currently going to these top-sectors, the funded projects can be considered as a useful representation of *research collaboration for innovation*. In order to investigate geographical properties of these collaboration networks we collected data about the projects, and statistical data about the characteristics of the container geographical units. These data are openly available on The Dutch data portal[30]. We deployed the following open datasets:

- *RVO dataset*[31] provides a list of R&D projects that have received subsidies and financial support from the Netherlands Enterprise Agency[32]. Projects information includes companies and research institutes which are collaborating on the project together with the geographical coordinates of the project partners.
- *CBS dataset*[33] published by the statistics office of the Netherlands[34] provides different types of statistical information on dimensions such as labor, income, economic activities, demography, and socio-cultural characteristics of the regions in the Netherlands.

As we did not know ex ante what the level of geographical organization of the consortia was, we needed to define these in different granularity. This enabled us to find out at what geo-level the consortia were organized. We could then identify the characteristics of these geographical 'containers' of the projects. To realize that, we first calculated different sets of UAs based on different statistics provided by the CBS dataset and different levels of open administrative boundaries. Figure 13 shows the delineation of these Urban Areas through population, business establishment, and combinations of these two indicators in the municipality level: *Boundaries typically differ when defined by different characteristics*.

When compared to the OECD FUAs[35] (right map in Figure 13), the adaptive UA approach enables the user to put different weights on the regional characteristics, leading to different (administrative) boundaries and to different patterns of regional organization.

Using the SMS geo-enrichment module we mapped geographical coordinates of RVO projects on these UAs to analyze the relation of the number projects with the selected socio-economic regional characteristics. Figure 14 shows the result of the mapping where frequency of the projects on different factors are highlighted: the darker the color, the higher the number of awarded projects. As can be seen when comparing Figure 13 and Figure 14, by far not all (F)UAs have projects. But more importantly, the different ways the UAs are defined leads to different outcomes. The OECD FUAs (right map), and the population density based UA (left map) miss some of the relevant areas[36]. And when one tries to identify the properties of innovative areas, using the FUA classification would sim-

[30] https://data.overheid.nl
[31] http://www.rvo.nl/open-data-van-rvonl
[32] Rijksdienst voor Ondernemend Nederland: RVO.nl
[33] https://www.cbs.nl/en-gb/our-services/open-data
[34] Het Centraal Bureau voor de Statistiek: CBS.nl
[35] We used the Eurostat shapefile that was used in the large 'urban audit' study (http://ec.europa.eu/eurostat/web/cities/data/database)
[36] in this case South-West Friesland is missing in population based indicators because it is less populated but still hosts a large set of businesses.

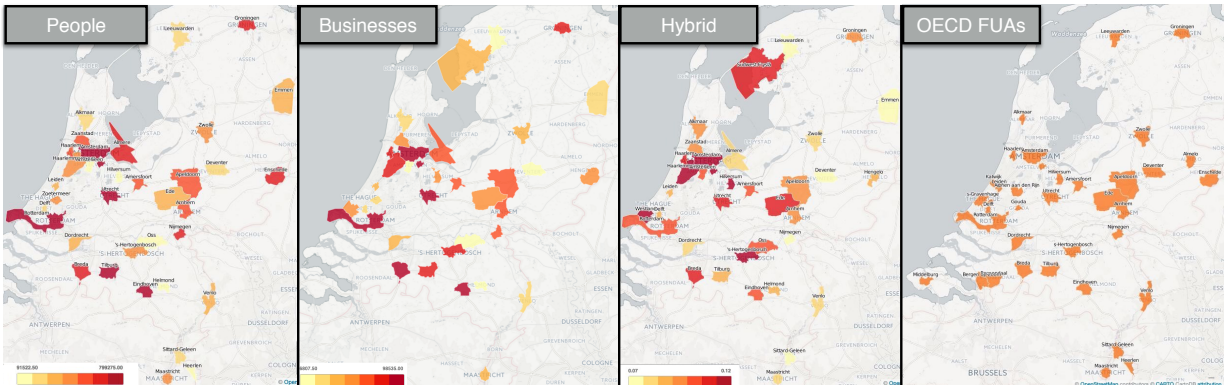

Fig. 13. An example of the adaptive delineation of FUAs for the Netherlands based on the open statistical data (populations, business establishments, hybrid and OECD).

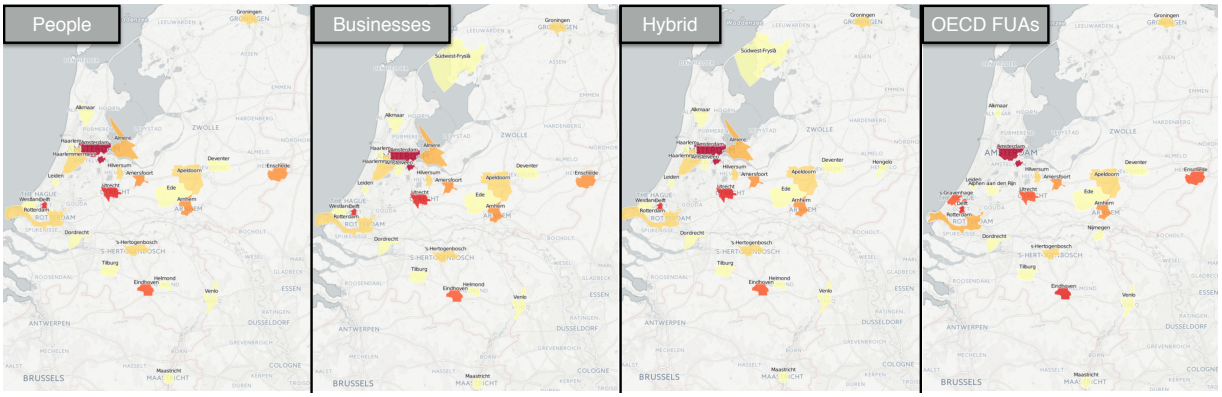

Fig. 14. Amount of RVO project subsidies mapped to the dynamically delineated FUAs defined based on the CBS open statistical data and OpenStreetMap boundaries.

ply miss innovative areas that do not fit in the FUA definition.

8.2. *Analyzing change in Higher Education systems*

The SMS data store contains many datasets with information about organizations. This use case is on detecting structural change in higher education (HE) systems. For this, the faceted browser was of great help, as it enables to explore the available information in a visual/graphical form. While browsing the datasets, the user found a property "foundation year". Selecting that property for a country immediately shows the frequency of new HE institutions per year. The browser showed an exceptional high concentration of new establishments in two consecutive years: In 1986 and 1987 some 21 new HE institutions were founded in the Netherlands, on a total of 114. So some substantial changes in the HE system seemed to have taken place (cf. Figure 16). By selecting these two years, a table with the names of these new organizations is shown. As the data were linked to other datasets, but also to their own website and their Wikipedia page, much more information on the newly established organizations was available. The system did not only provided many numerical data about the organizations, such as numbers of students and staff, but also qualitative (textual) data. Inspecting this information showed that all these new organizations are Universities of Applied Sciences, which is the "second tier" Dutch higher education. By reading the historical information on their Websites, one would find out that the new (often very large) institutions in fact are mergers of regional conglomerations of smaller schools, a major reform of the Dutch HE system.

A follow-up question was whether this was a typical Dutch phenomenon, or whether similar changes have taken place in other countries. Belgium was the second

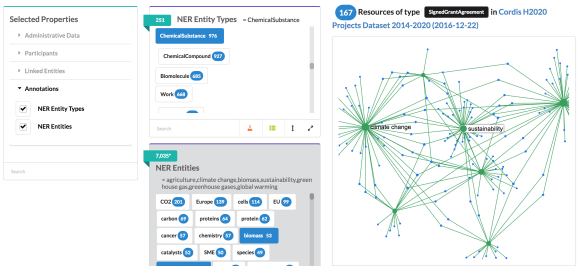

Fig. 15. Deploying background knowledge to facilitate browsing of data.

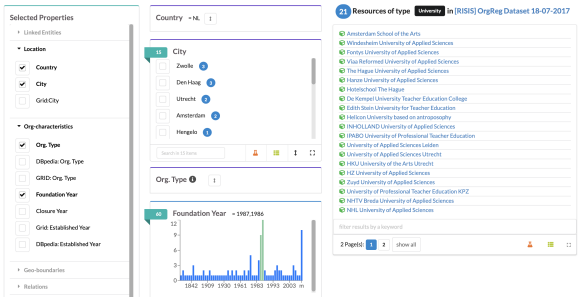

Fig. 16. Analyzing change in Higher Education systems in the Netherlands.

case to inspect, following the same steps. Also here the browser showed concentrations of foundations of new HE institutions, but now in the year 1995 when 32 new HE institutions were founded in Belgium. Selected the year 1995 provided a list with the names of the newly founded institutions. Inspecting the list of names showed (even without prior knowledge) that the changes probably took place in the French speaking part of Belgium and not in the Flemish speaking part, as all institution names are French. Inspecting the further available information then may lead to a further clarification of the changes that took place.

The third case was Austria, and also there a concentration of new institutions was detected - in 2007, so a decade after the changes in Belgium and two decades after the changes in the Netherlands. Of the total of 102 HE institutions in Austria, 15 were created in 2007 - again a percentage suggesting some form of structural change. Selecting the entity type in the browser shows that the changes have taken place in the sector of teacher education: the newly founded HE institutions are all of type "University of Education", "University College of Teacher Education", and "Pedagogical University". Obviously, the changes in the Austrian system were less broad than in the Netherlands or in Belgium, where the changes seem to cover a much larger part of the HE system. Inspecting the other linked information - qualitative and quantitative - was very helpful for a deeper study or what took place and where.

8.3. Explaining the ranking of universities

University rankings, such as the THE ranking and the Leiden Ranking have become important indicators for the quality of universities. But what explains the position of a university on a ranking? Of course, one may think of characteristics like the student/staff ratio, or the total research budget. But also contextual variables may be relevant, such as the number of research intensive organizations in the environment of a university, as this may create opportunities for funding and collaboration, and it may attract students because of the labor market conditions. Also a vibrant local environment may be relevant, as this may attract high quality staff. Answering the question requires data from different sources, which therefore need to be linked. The following datasets contain information needed:

University rankings: The Leiden Ranking dataset offers scientific performance indicators of more than 2500 universities above the threshold of 100 fractionally counted Web of Science indexed core publications. The threshold explains why it does not cover all countries.

Characteristics of the university: ETER, the European Tertiary Education Register is a database on European Higher Education Institutions, covering 35 countries (in the 2015 version). Several types of HE institutions are distinguished, such as research universities, professional schools, etc. As it provides statistics on European Universities, it is helpful to answer the research question for European universities.

Research intensive organizations: The Global Research Identifier Database (GRID - version May 2018) describes 87868 organizations across 217 countries using 14220 relationships. Some seventeen countries, among those the United States, United Kingdom, Japan, Germany, France, Canada, China, India, Italy, Spain, Brazil, and Russia, have more than a thousand organizations, accounting for 77% of the total. All organizations within GRID are assigned an address, while 96% of them have an organization type (company, education, health care, non-profit, facility, government, archive, and 'other'). Furthermore, 80% of the organizations have geographic coordinates. We use GRID for counting the number (and type) of organizations that surround the universities.

Geographical boundaries: Neither ETER nor GRID define geographical boundaries. For this, GADM, the Database of Global Administrative Areas, is deployed. It provides maps and spatial data for all countries and their regions, provinces, municipalities and neighborhoods. Since GADM provides geographical boundaries and since both ETER and GRID are enriched with coordinates, GADM could be used to add to ETER and GRID the GADM level-2 geographical boundary code. Figure 17 illustrates highlights the enrichment of the two datasets using GADM.

Local and regional data: For the other characteristics of the environment of the universities - such as on demographics, income, cultural facilities - one can use open data from the national statistical offices. In order to use the five datasets described above, *data integration* was done according to the integration model depicted in Figure 17. It is based on the Lenticular Lens approach for connecting entities from various heterogeneous datasets as described in the section on 'Linking earlier' in this paper. The implemented integration data model is shown in Figure 17, and Figure 18 shows a Lenticular Lens composed of six linksets and two lenses.

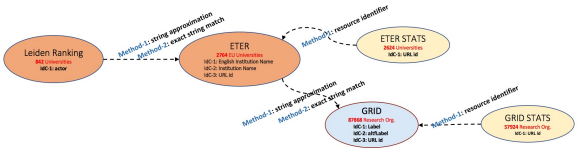

Fig. 17. Contextual Integration using the Lenticular Lens approach.

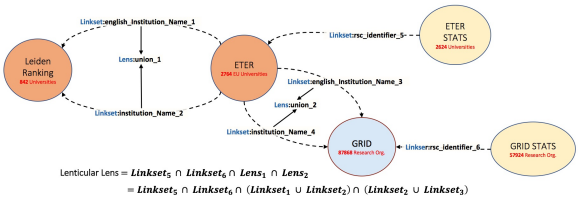

Fig. 18. Contextual Integration using the Lenticular Lens approach.

Linksets 1 and 2 were used to establish identity links between Leiden ranking and ETER. These are based on either the English or local name of organizations in ETER organizations, and on the actor names in the Leiden Ranking, using an exact string match algorithm. The same identity link discovery process was executed between ETER (English or local name) and GRID (label or altLabel) resulting in *linkset 2 and 3*. *Lens 1* enables putting together all links found between Leiden and ETER while *lens 2* does the same between ETER and GRID. We used *linkset 5 and 6* to restrict the links to those for which a geographical boundary was found in GADM. The outcome of this integration makes it possible to retrieve the attributes needed from different datasets. This can be done either through the faceted browser, or by writing a query.

In order to test the approach, we applied it on the research universities in the Netherlands. Figure 19 gives an example of the output. The table contains data from ETER (columns 1 to 3, 8,9, 10), from ETER enriched with the geo-location tool as described earlier in this paper (columns 4 and 5), from GADM (column 6), from GRID and GADM (column 7), and from the Leiden Ranking (column 11).

Correlating (with the Dutch universities only) the ranking score (column 11) with the number of R&D intensive organizations (column 7) gives a correlation of 0.58. The result of this small scale test suggest that further analysis is worthwhile in order to better understand what factors explain university rankings.

8.4. Evaluating research portfolios

Rankings may be important, it is as important to be able to describe and compare research portfolios of universities and research institutes. This is even more the case now the societal and economic impact of research have become a focal concern in science and innovation policy. To what extent do research portfolios cover the important scientific challenges and the important societal challenges, as e.g., formulated in the UN millennium development goals? The same can be asked about research funding organizations: Is the portfolio of granted projects adequate in terms of the societal challenges?

In this use case, we deploy the SMS annotation tool and the CORDIS dataset to explore these issues. The CORDIS dataset contains several relevant characteristics of the projects, such as organizations involved, the organization type, the program the project belongs to, the start and end year, budgets, and a text summarizing the project. Using the SMS annotator, the project summaries were annotated with the terms from the DBpedia ontology. It includes more technical terms, but also many general encyclopedic concepts. Adding these extracted terms as descriptors of the project has a great advantage, as we can combine technical research terms and the more general policy related terms to retrieve the relevant projects. The annotation enables the user to browse data using two new facets, one for extracted

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. 19. Use case 8.3: part of the resulting dataset (Dutch universities only, and a few of the variables)

terms and one for the taxonomy these terms belong to (cf. Figure 15).

In this way, the annotator helps to solve the problem of finding what research projects are relevant for specific societal challenges, a core problem in assessing research programs. As an example, one may want to know what chemistry projects were granted that focus on sustainability issues. Using the 'entity types' from the annotator, the class 'chemical substances' was selected. This results in some 1300 funded projects in H2020 up to 2018. Then one can select - using the more than 8000 'entities' - all categories that relate to sustainability, such as sustainability, carbon dioxide, greenhouse gas, climate change, biodegradable, solar cells, and so on. This leads to some 500 projects, suggesting that some 40% of the funded chemistry projects are related to sustainability issues.

Having selected the relevant projects, one can then further analyze this specific project portfolio in different dimensions. Questions can be answered like: Is attention for a specific societal challenge increasing or decreasing? What is the level of multidisciplinarity of the granted projects? Who are involved in the projects (countries, universities, fields)?[37] Because the resulting set for a very specific topic is generally not too large, the user may even manually inspect the policy-science link.

[37] More details are presented in [11, 23].

9. Related Work

Similar work has been done within other domains. In the pharmaceutical domain, the OpenPHACTS project[38] brings together pharmacological data resources in an integrated and interoperable infrastructure. For example, it connects information about chemistry to biological information such that researchers can determine the potential impact of a chemical on a biological system [12]. In the humanities domain, CLARIAH[39], the Common Lab Research Infrastructure for the Arts and Humanities, is another digital infrastructure project that also brings together large collections of data and software from different humanities disciplines. In the same domain, CEDAR (Census RDF Data) interlinks Dutch census data with other hubs of historical socio-economic data, demographic data and more, to create a semantic data-web of historical information. This allows researchers to bridge the diversity of sources of information [24] for historical research, and enables to ask complex questions that involve among others historical data and socio-economic data. Covering a much broader set of scientific disciplines, the Center for Expanded Data Annotation and Retrieval also known as CEDAR[40] provides a unified framework for creating consistent and easily searchable metadata in all scientific domains.

[38] https://www.openphacts.org/
[39] http://www.cltl.nl/projects/current-projects/clariah/
[40] https://med.stanford.edu/cedar/our-solution.html

The OpenAIRE [41] and EUDAT[42] digital data infrastructures together with the European Open Science Cloud (EOSC) initiative are related infrastructures. OpenAIRE offers access to open publications, including quite some metadata, but less functionality to link it to other data. That is why the SMS platform has included the OpenAIRE datasets. EUDAT is an umbrella system that offer European researchers and science and technology professionals a virtual environment to access a variety of existing data infrastructures.

10. Conclusions and Future Work

Progress in the social sciences, in our opinion, depends on the ability to perform 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, and this requires high quality linked and enriched data under FAIR principles, that can be accessed through user-friendly interfaces. The SMS platform addresses this need by providing a data handling and analysis infrastructure to meet the needs of researchers, policy makers, and managers in the area of science, technology, and innovation studies. Its ultimate goal is to support a variety of users aiming to investigate and improve the science, innovation and higher education system in a better way than what has been possible heretofore – by exploiting the power that comes from linking big and heterogeneous data and by putting the emphasis on open access to research data.

Most of the services and the user interfaces provided by the SMS platform are generic and domain-independent. We are convinced that the SMS infrastructure can be used for data linking, enrichment and analysis in other social science domains and beyond. In fact, SMS is already used by researchers from geography, public health, and educational research.

For the future we envisage to work on the following issues and improvements:

a) Adding more data from the Web, from social media and other streaming data sources.

b) Adding support for adaptive annotations that allow using arbitrary knowledge graphs for annotating datasets. This feature would enable search and browsing of data using a more deep and specific knowledge rather than generic knowledge from Wikipedia.

c) Adding more emphasis on data quality assessment to make sure that the quality of results generated on the platform reach a good level of quality.

d) Add support for automatic machine learning techniques that would further facilitate the data analysis on the SMS platform.

e) Sharing, integrating, and managing resources across infrastructures is crucial for any sustainable business plan. As future work, two levels of interoperability and sharing are foreseen: (i) Harmonizing SMS service interfaces and aligning them with those of other EU and global research infrastructures; (ii) Realizing together with other infrastructures an integrated hub for joint applications based on linked open data, FAIR, and smartAPI principles[43].

[41] https://www.openaire.eu

[42] https://eudat.eu

[43] https://smart-api.info

References

[1] E.J. Hackett, O. Amsterdamska, M. Lynch and J. Wajcman, *The handbook of science and technology studies*, The MIT Press, 2008.

[2] A. Khalili, P. van den Besselaar, A.K. Idrissou, K.A. de Graaf and F. van Harmelen, Semantically Mapping Science (SMS) Platform, in: *Proceedings of the First Workshop on Enabling Open Semantic Science co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21st, 2017.*, D. Garijo, W.R. van Hage, T. Kauppinen, T. Kuhn and J. Zhao, eds, CEUR Workshop Proceedings, Vol. 1931, CEUR-WS.org, 2017, pp. 1–6. http://ceur-ws.org/Vol-1931/paper-01.pdf.

[3] A.K. Idrissou, A. Khalili, R. Hoekstra and P. van den Besselaar, Managing metadata for science, technology and innovation studies: The RISIS case, in: *WHiSe*, A. Adamou, E. Daga and L. Isaksen, eds, CEUR Workshop Proceedings, Vol. 1608, CEUR-WS.org, 2016, pp. 15–20. http://ceur-ws.org/Vol-1608/paper-03.pdf.

[4] A. Khalili, P. van den Besselaar and K.A. de Graaf, Using Linked Open Geo Boundaries for Adaptive Delineation of Functional Urban Areas, in: *The Semantic Web: ESWC 2018 Satellite Events - ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3-7, 2018, Revised Selected Papers*, A. Gangemi, A.L. Gentile, A.G. Nuzzolese, S. Rudolph, M. Maleshkova, H. Paulheim, J.Z. Pan and M. Alam, eds, Lecture Notes in Computer Science, Vol. 11155, Springer, 2018, pp. 327–341. doi:10.1007/978-3-319-98192-5_51. https://doi.org/10.1007/978-3-319-98192-5_51.

[5] A.K. Idrissou, R. Hoekstra, F. van Harmelen, A. Khalili and P. van den Besselaar, Is my: sameAs the same as your: sameAs?: Lenticular Lenses for Context-Specific Identity, in: *Proceedings of the Knowledge Capture Conference*, ACM, 2017, p. 23.

[6] A.K. Idrissou, F. van Harmelen and P. van den Besselaar, Network Metrics for Assessing the Quality of Entity Resolution Between Multiple Datasets, in: *Knowledge Engineering and Knowledge Management*, Accepted paper, EKAW 2018, Nancy, France.

[7] A. Khalili, A. Loizou and F. van Harmelen, Adaptive Linked Data-Driven Web Components: Building Flexible and Reusable Semantic Web Interfaces, in: *ESWC*, Lecture Notes in Computer Science, Vol. 9678, Springer, 2016, pp. 677–692. doi:10.1007/978-3-319-34129-3_41. https://doi.org/10.1007/978-3-319-34129-3_41.

[8] A. Khalili, P. van Andel, P. van den Besselaar and K.A. de Graaf, Fostering Serendipitous Knowledge Discovery using an Adaptive Multigraph-based Faceted Browser, in: *Proceedings of the Knowledge Capture Conference, K-CAP 2017, Austin, TX, USA, December 4-6, 2017*, Ó. Corcho, K. Janowicz, G. Rizzo, I. Tiddi and D. Garijo, eds, ACM, 2017, pp. 15–1154. doi:10.1145/3148011.3148037. http://doi.acm.org/10.1145/3148011.3148037.

[9] A. Khalili and K.A. de Graaf, Linked Data Reactor: Towards Data-aware User Interfaces, in: *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, R. Hoekstra, C. Faron-Zucker, T. Pellegrini and V. de Boer, eds, ACM, 2017, pp. 168–172. doi:10.1145/3132218.3132231. http://doi.acm.org/10.1145/3132218.3132231.

[10] A. Khalili and A. Meroño-Peñuela, WYSIWYQ - What You See Is What You Query, in: *Proceedings of the Third International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 22, 2017.*, V. Ivanova, P. Lambrix, S. Lohmann and C. Pesquita, eds, CEUR Workshop Proceedings, Vol. 1947, CEUR-WS.org, 2017, pp. 123–130. http://ceur-ws.org/Vol-1947/paper11.pdf.

[11] A. Khalili, P. van den Besselaar and K.A. de Graaf, FERASAT: A Serendipity-Fostering Faceted Browser for Linked Data, in: *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, A. Gangemi, R. Navigli, M. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai and M. Alam, eds, Lecture Notes in Computer Science, Vol. 10843, Springer, 2018, pp. 351–366. doi:10.1007/978-3-319-93417-4_23. https://doi.org/10.1007/978-3-319-93417-4_23.

[12] P. Groth, 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.

[13] R.C. Amorim, J.A. Castro, J.R. da Silva and C. Ribeiro, A comparative study of platforms for research data management: interoperability, metadata capabilities and integration potential, in: *New contributions in information systems and technologies*, Springer, 2015, pp. 101–111.

[14] W. Beek, S. Schlobach and F. van Harmelen, A Contextualised Semantics for owl: sameAs, in: *Extended Semantic Web Conference (ESWC)*, Springer, 2016, pp. 405–419.

[15] J. Raad, W. Beek, F. van Harmelen, N. Pernelle and F. Saïs, Detecting Erroneous Identity Links on the Web Using Network Metrics, in: *International Semantic Web Conference*, Springer, 2018, pp. 391–407.

[16] S. Auer, J. Lehmann, A.-C. Ngonga Ngomo and A. Zaveri, Introduction to Linked Data and Its Lifecycle on the Web, in: *Proceedings of the 9th International Conference on Reasoning Web: Semantic Technologies for Intelligent Data Access*, RW'13, Springer-Verlag, Berlin, Heidelberg, 2013, pp. 1–90. ISBN 978-3-642-39783-7. doi:10.1007/978-3-642-39784-4_1. http://dx.doi.org/10.1007/978-3-642-39784-4_1.

[17] P. van Andel, Anatomy of the Unsought Finding. Serendipity: Origin, History, Domains, Traditions, Appearances, Patterns and Programmability, *Br J Philos Sci* **45**(2) (1994), 631–648. doi:10.1093/bjps/45.2.631. http://dx.doi.org/10.1093/bjps/45.2.631.

[18] J. Hendler and A. Hugill, The syzygy surfer:(Ab) using the semantic web to inspire creativity, *International Journal of Creative Computing* **1**(1) (2013), 20–34.

[19] A. Acosta, Using serendipity to advance knowledge building activities, *Ontario Institute for Studies in Education, University of Toronto, Canada* (2012).

[20] L. Yu, Follow Your Nose: A Basic Semantic Web Agent, in: *A Developer's Guide to the Semantic Web*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014, pp. 711–736. ISBN 978-3-662-43796-4. doi:10.1007/978-3-662-43796-4_16. http://dx.doi.org/10.1007/978-3-662-43796-4_16.

[21] E.G. Toms et al., Serendipitous Information Retrieval, in: *DELOS Workshop*, 2000, pp. 17–20.

[22] OECD, *Redefining "Urban": A New Way to Measure-Metropolitan Areas*, OECD, 2012. ISBN 9789264174054. doi:10.1787/9789264174108-en.

[23] P. van den Besselaar, A. Khalili and U. SandstrÃűm, Evaluating research portfolio's through ontology based text annotation, in: *the 7th Global TechMining Conference (GTM) 2017*, 2017. http://www.gtmconference.org.

[24] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach and F. Van Harmelen, Semantic technologies for historical research: A survey, *Semantic Web* **6**(6) (2014), 539–564.