Managing metadata for science, technology and innovation studies: The RISIS case

Al Koudous Idrissou[1,3], Ali Khalili[1], Rinke Hoekstra[1,2], and Peter van den Besselaar[3]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, NL
{o.a.k.idrissou,a.khalili,rinke.hoekstra}@vu.nl
[2] Faculty of Law, University of Amsterdam, NL
[3] Department of Organization Sciences, Vrije Universiteit Amsterdm, NL
p.a.a.vanden.besselaar@vu.nl

Abstract. Here, we describe the RISIS-SMS metadata system, developed to support the use of heterogeneous datasets in the field of Science, Technology and Innovation Studies (STIS). These data are partly within the RISIS infrastructure, but often elsewhere. The system has three aims: (i) to help researchers to search for and understand data that will help to answer specific research questions, without having to access or download the data. As datasets often have restricted access, browsing metadata is a key feature of the system: researchers need help identifying the relevant data from different sources for their research, and for which data it is worthwhile asking for access; (ii) to support the enrichment of data By linking the metadata system to the Linked Open Data environment (LOD); (iii) to facilitate application-driven data integration.

Keywords: metadata, Linked Data, Science & Technology Studies, Research Infrastructures, digital humanities

1 Introduction

The field of Science, Technology and Innovation Studies (STIS) is an interdisciplinary field between the social sciences and the humanities. It covers many fields from the economics of science and innovation up to the history and philosophy of science [5]. It relies on the availability of a large volume of highly heterogeneous data: *structured* and *unstructured*, *qualitative* and *quantitative*. STIS studies the *dynamics of scientific ideas* by analysing the content of scientific publications and project descriptions. For example, to help understand the *selection processes* taking place in the scientific community or to better understand *life histories* of scientists and research organizations.

Based on requirements extracted from interviews we conducted, we identify the need for researchers to search across datasets and for data providers to attract researchers while keeping restricted datasets access. To address these problems, we describe in this paper a *rich RDF metadata vocabulary* to overcome access limitations and facilitate data *discovery*, *integration* and *use* by humans

and machines [3]. The vocabulary is used by the Semantically-Mapping Science[4] infrastructure (SMS) to provide metadata services besides Geo services, Integration services and Category services. It enables the integration of qualitative and quantitative approaches that have been strongly diverging over time [1].

2 Requirements

Problem Summary. At its core, SMS comprises a collection of proprietary and public databases relevant to the field of STIS. For most RISIS datasets, access is restricted to users with authorization only granted after an explicit request is submitted to the data owner. Other data can only be accessed on site at a *physical location*. This gets in the way of a good understanding of the content and coverage of a dataset: if data is not reachable, how can one decide if a closed dataset is relevant enough for addressing her research question before requesting for access, or travel to visit a dataset owner? This *access limitation* does not only hinder findability and relevance assessment ex ante, but also hinders the ex post integration of heterogeneous datasets after they have been identified.

Methodology. To address this problem we designed a metadata vocabulary guided by informal interviews conducted with STIS researchers and data-owners.This helped to identify and categorize (Figure 1) the information the metadata should cover. After the first version was developed, owners of some 12 datasets used the system. We visited all the data-owners, and discussed the user experiences as well as the benefits and problems. This was used to improve the vocabulary design. Finally exchanging email with users helped in fine tuning the metadata.

Interviews Outcome & Solutions. ***Protect proprietary datasets*** - To protect datasets that contain private and sensitive data or data for which a specific permission or subscription is required, we categorised RISIS datasets as: *confidential* data, *publicly accessible* data, and all other relevant *public* data on the Web. ***Overcome access limitation*** - Because data access is limited, we provide users with means for browsing dataset metadata rather than inspecting the data itself. The metadata should meet six requirements: ***R1*** - Facilitate information *displaying* at a user interface level. ***R2*** - Provide information *guiding the use of data*. ***R3*** - Provide *detailed information* about the datasets available on RISIS-SMS. ***R4*** - Support users to get an *in-depth understanding* of the data at hand, in such a way that they easily identify how the data should be interpreted, used, or linked to other data. ***R5*** - Facilitate trust by providing details about the *quality* of the underlying data. ***R6*** - Facilitate *simple and advanced search* for relevant data. The latter is considered to be a crucial task for data discovery and link discovery across datasets. ***Trigger research opportunities*** - Use LOD to organize and integrate databases far beyond the internal RISIS datasets to create new opportunities for research. For example a link to a city or a university described in DBpedia or GeoNames could be exploited to infer new knowledge such as the entity location or geographical boundaries.

[4] more details are available at http://sms.risis.eu

Metadata Operationalization. From the interviews, we concluded that our metadata should cover a broad range of different aspects which we categorised into the following nine metadata types: dataset details {overview, temporal aspects, content, structure, technical aspects, legal aspects, access, visit and data quality/used methodologies} and person details. Detailed description about each of the aforementioned aspects is beyond the scope of this paper. However, we shortly describe here how the categories are mapped to the requirements. *Technical aspects*, *access-visit* which provides information on the *conditions for visiting* one or more data-owners and *legal* aspects, this satisfies ***R2***. Data *overview*, *description* of the *content*, *temporal aspects* and *structure of the data*, this supports ***R3*** and ***R4***. The *provenance* which helps to know the *origin*, *creator*, *when* and *how* of the data, this satisfies ***R5-Trust***. The *methodologies* followed for addressing some dimensions of *data quality* such as records de-duplication, resource disambiguation and, data consistency and correctness, this again satisfies ***R5-Trust***. All aspects of the metadata could be used for simple search. For complex search, only attributes that link to external knowledge are exploitable ***R6***. ***R1*** does not follow the above mapping. Instead, it is covered by the creation of a User-friendly Interface[5] that addresses *Categorization* of metadata types, use of *non technical words* and text hint.

3 Reused Vocabularies

Since the RISIS metadata covers a broad range of aspects, a platform or a vocabulary that is all inclusive does not exist. So, we selected a set of nine domain-specific dataset metadata vocabularies designed for one or more aspects discussed in the requirements. The main vocabularies identified to share many terms within RISIS metadata concepts are respectively DCMI,[6] PROV-O,[7] VoID[8] and FOAF.[9] Although provenance is not shown in Figure 1, we use it extensively behind the scene for describing data manipulations. Other reused vocabularies that involved less coverage of the RISIS requirements include DCAT,[10] DISCO,[11] WAIVER,[12] PAV [2] and SKOS.[13] Figure 1 illustrates the mapping between requirements and existing shared vocabularies where, the table header expresses a domain, the namespace prefix indicates the vocabulary and the suffix (local name) indicates the term mapped to a requirement term. Yet, reusing these vocabularies did not entirely satisfy RISIS's view on describing a dataset. As a result, we created new terms such as `risis:usecase` (see Figure 1) for concepts that were not covered by any of the selected vocabularies.

[5] User-friendly Interface designed for the matter http://datasets.risis.eu/
[6] Dublin Core Metadata Initiative: http://dublincore.org/documents/dcmi-terms/
[7] PROV Ontology: https://www.w3.org/TR/2013/REC-prov-o-20130430/
[8] Vocabulary of Interlinked Datasets: https://www.w3.org/TR/void/
[9] FOAF Vocabulary: http://xmlns.com/foaf/spec/
[10] Data Catalog Vocabulary: https://www.w3.org/TR/vocab-dcat/
[11] Disco Vocabulary: http://rdf-vocabulary.ddialliance.org/discovery.html
[12] Waivers of rights vocabulary: http://vocab.org/waiver/terms
[13] SKOS Vocabulary: https://www.w3.org/TR/swbp-skos-core-spec

The work done so far is a pilot that needs a more exhaustive investigation of six other potential vocabularies (see footnote) before finalizing the RISIS ontology. A better understanding of these vocabularies is expected lead to (1) in depth understanding for better data integration at the schema level;[14] (1,2) facilitate publishing mapping between a tabular dataset and its RDF converted version;[15] (3) offer providers and consumers means to assess the quality of datasets;[16] (4,5) using the right concept describing statistical information relevant for RISIS and, for publishing RISIS multi-dimensional statistical data;[17,18] and (6) to applying the concept of "One ontology to bind them all" [7] to the RISIS problem and better coverage of legal aspects such as licensing.[19]

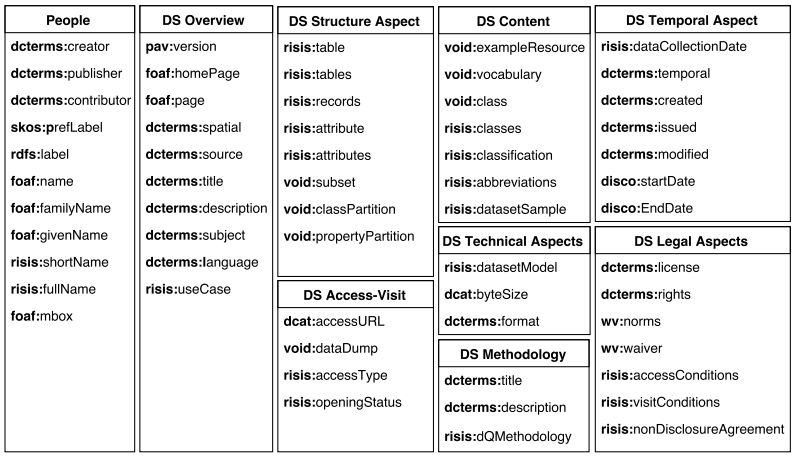

Fig. 1. The RISIS Ontology. A view over mapped vocabularies reused. The table header expresses a domain, the namespace prefix indicates the vocabulary and the suffix (local name) indicates the term mapped to a requirement term.

4 Related Work and Discussion

Related work. ***Projects*** - Open PHACTS[20] gathers pharmacological resources in an integrated and interoperable infrastructure to connect for example, information about chemistry to biological information such that users can determine the potential impact of a chemical on a biological system [4]. CLARIAH[21] also gathers large collections of data and software from different humanities disciplines.

[14] Vocabulary for Tabular Data: https://www.w3.org/TR/tabular-metadata/
[15] D2RQ Mapping Language: http://d2rq.org/d2rq-language
[16] Data Quality Vocabulary: https://www.w3.org/TR/vocab-dqv/
[17] SDMX vocabulary: http://lov.okfn.org/dataset/lov/vocabs/sdmx
[18] Data Cube vocabulary: https://www.w3.org/TR/vocab-data-cube/
[19] Meta-Share OWL meta model: http://purl.org/ms-lod/MetaShare.ttl
[20] The Open PHACTS project: https://www.openphacts.org/
[21] CLARIAH: http://www.clariah.nl

CEDAR[22] inter-links Dutch census data with other data hubs to create a semantic data-web of historical information. Such construct allows researchers to bridge information diversity [8] for historical research to ask complex questions that involves historical, socio-economic, demographic data and more domains. The Center for Expanded Data Annotation and Retrieval also known as CEDAR[23] provides a unified framework for researchers in all scientific disciplines who need to create consistent and easily searchable metadata. ***Platforms*** - Datahub.io provides a public registry for datasets and metadata-based data discovery. Only, its metadata coverage is not adequate for, the language, provenance and license of a resource are not properly represented. http://lodlaundromat.org/ gives free access to LOD collected on the web. Although it produces valuable information for data comparison, analytics and more, it fails to provide sufficient description satisfying R1-6. ***Vocabularies*** - DCAT does not provide format specific information, it does not provide information on the content of a dataset nor does it describe the provenance of data and, its coverage of legal aspects of a dataset is limited. Open PHACTS uses VoID, a data specific vocabulary[24]. Through usage, the specifications defined in there turned out to be so strict that, creating a proper VoID document was too hard to manage by developers; something RISIS planed to avoid with its metadata. ***Summary*** - Projects that gather data from various sources and domains exist in disciplines such as Pharmaceutical, Art, Humanities, Socio-economic, Historical, etc. but not in STIS. Platforms that provide dataset metadata also exist. Only, they are limited or too specific. Although many shared vocabularies exit, they are not rich enough.

Discussion. Access limitations and related work described in this paper underpin the increasing demand for bringing together data from multiple domains to allow for complex and cross-domain analysis. We argue that this would not be done without the creation of metadata for systematic and consistent description of collections of datasets within a flexible data model. The need for a User-friendly Interface[25] (UI) arose from choosing RDF as the metadata model for, it facilitates integration and information sharing on the web. In fact, RISIS data-owners who need to generate and maintain metadata about their datasets are not familiar with the Semantic Web technologies and the ways to generate a standard and valid RDF. As a result, following a user-centred approach, the proposed vocabulary was implemented as a UI to help data-owners to easily auto-generate and manipulate RDF metadata based on the RISIS metadata vocabulary. The UI has been in use by RISIS data providers. Exposing the metadata through the RISIS User-friendly Interface [6] stimulated data providers to check the quality of their data before opening it for access/visit. Securing the data quality in terms of the standards agreed upon in RISIS was done by satisfying elements in a RISIS defined check-list before the opening of the data. Given the broad scope

[22] Census data open linked: https://www.cedar-project.nl/
[23] CEDAR: https://med.stanford.edu/cedar/our-solution.html
[24] http://www.openphacts.org/specs/2013/WD-datadesc-20130912/
[25] An example of the RISIS UI http://datasets.risis.eu/metadata/eupro

and generic domain of the problem, SMS is intended to be useful not only for STIS but also for the humanities and social sciences.

5 Conclusions and Future Work

This paper presents an approach for managing metadata in the field of science, technology and innovation studies. The approach was developed and applied in the context of the RISIS-SMS project with the goal of *supporting data integration, discovery* and *search across datasets*, *maintaining privacy*, and *obtaining user trust* while *focusing on data that are not directly accessible*. A contribution of this work is the *requirements* elicited by *interviewing the stakeholders*. The requirement analysis guided the *design of a new vocabulary*, together with *review of existing metadata vocabularies* that helped us filling in part of the metadata needed to accommodate the domain needs. Additionally, to meet the requirements, we designed and implemented a *user-friendly interface* which allows non-expert users to easily author metadata in RDF.

As future work, we envisage to extend our vocabulary to cover aspects related to the quality and provenance of data. We also plan to conduct a usability evaluation with end-users of the system to ensure that our user interface and metadata specifications fulfil the user needs.

References

1. Van den Besselaar, P.: The cognitive and the social structure of sts. Scientometrics 51(2), 441–460 (2001)
2. Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J., Goble, C., Clark, T.: PAV ontology: Provenance, authoring and versioning. Journal of biomedical semantics 4(1), 1–22 (2013)
3. Daraio, C., Lenzerini, M., Leporelli, C., Moed, H.F., Naggar, P., Bonaccorsi, A., Bartolucci, A.: Data integration for research and innovation policy: an ontology-based data management approach. Scientometrics pp. 1–15 (2015)
4. Groth, P., Loizou, A., Gray, A.J., Goble, C., Harland, L., Pettifer, S.: API-centric linked data integration: The Open PHACTS discovery platform case study. Web Semantics: Science, Services and Agents on the World Wide Web 29, 12–18 (2014)
5. Hackett, E.J., Amsterdamska, O., Lynch, M., Wajcman, J.: The handbook of science and technology studies. The MIT Press (2008)
6. Khalili, A., Loizou, A., van Harmelen., F.: Adaptive linked data-driven Web components: Building flexible and reusable Semantic Web interfaces. Semantic Web Conference (ESWC) 2016 (2016), to appear
7. McCrae, J.P., Labropoulou, P., Gracia, J., Villegas, M., Rodríguez-Doncel, V., Cimiano, P.: One ontology to bind them all: The META-SHARE OWL ontology for the interoperability of linguistic datasets on the Web. In: The Semantic Web: ESWC 2015 Satellite Events, pp. 271–282. Springer (2015)
8. Meroño-Peñuela, A., Ashkpour, A., Van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., Van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web 6(6), 539–564 (2014)