A platform for scientific data sharing

Karima Rafes, Cécile Germain

A platform for scientific data sharing

Une plateforme pour le partage de données scientifiques

Karima Rafes
BorderCloud & INRIA
karima.rafes@bordercloud.com

Cécile Germain
Université Paris Sud
cecile.germain@lri.fr

ABSTRACT

In this paper, we use the semantic web technology, notably RDF, SPARQL and Linked Open Data in the context of scientific data sharing. More precisely, we present the LinkedWiki platform that is being developed at the Center for Data Science of Paris-Saclay University for scientific data integration. The goal is to facilitate the discovery and exploitation of scientists' datasets by their colleagues. For this, we notably rely on the use by scientists of Wikipedia for specifying the semantics of datasets, and the use of Wikidata (the Wikipedia's knowledge base) identifiers for annotating these datasets and thereby facilitating their discovery.

RESUME

Dans cet article, nous utilisons la technologie du web sémantique, notamment RDF, SPARQL et Linked Open Data dans le cadre du partage de données scientifiques. Plus précisément, nous présentons la plate-forme LinkedWiki qui est développée au Center for Data Science de l'Université Paris-Saclay pour l'intégration de données scientifiques. L'objectif est de faciliter la découverte et l'exploitation de bases de données scientifiques par leurs collègues. Pour cela, nous nous appuyons notamment sur l'utilisation par les scientifiques de Wikipedia pour spécifier la sémantique des corpus de données et l'utilisation des identifiants Wikidata (la base de connaissance de Wikipedia) pour annoter ces ensembles de données et en faciliter ainsi la découverte.

Categories and Subject Descriptors

H.3.5 [**Information storage and retrieval**]: Online Information Services—*Data sharing*

; H.3.2 [**Information storage and retrieval**]: Information Storage—*Record classification*

; H.2.8 [**Database management**]: Database Applications—*Scientific databases*

BDA 29 septembre 2015, Ile de Porquerolles, France.

General Terms

Experimentation, Standardization, Management

Keywords

Linked Data, Open Science, Wikidata, Wikipedia, RDF, SPARQL

1. INTRODUCTION

The semantic Web comes with the promise to facilitate data sharing and reusing by scientists In this paper, we present a platform that takes advantage of knowledge management to ease the sharing of data by scientists and the linkage of data for facilitating the reconciliation between different ontologies.

With the LinkedWiki Platform[1], the scientist who publishes a data set specifies essential metadata. A key aspect is *perennial URIs* that are attached to concepts, and more precisely to concepts that are well-accepted on the Web. LinkedWiki uses the Linked Data technology to relate data sets to Web concepts, and the Wikidata ontology to capture concepts.

The paper is organized as follows. Section 2 discusses the context, in particular how scientific data workflows are used in scientific labs (simply called labs further) and problems that are encountered. Section 3 summarizes the current platforms that help scientists identify and locate online data repositories. Section 4 describes the main aspects of the LinkedWiki Platform. Section 5 describes the platform we have implemented. Finally, Section 6 is a conclusion.

2. THE CONTEXT

In this section, we first describe new operational issues that question reproducibility in the context of increasing data volume and quality requirements. Then we briefly discuss related basic technologies that can help meet these new challenges[2] .

2.1 New challenges with scentific data

With respect to data management, not all scientists are born equal. Large international collaborations organized around heavy instruments, e.g., High Energy Physics or

[1]An exemple is being deployed in the Center for Data Science (CDS) of Paris-Saclay University at `https://io.datascience-paris-saclay.fr`

[2]Due to space limitations, the bibliographic references to the well-known standards, products and platforms are not provided

Astrophysics, collectively define and enforce comprehensive data management and analysis processes, and devote considerable manpower to implement them with state of the art technology. But the rest of experimental science struggles with a permanent dilemma, how to make the data usable at minimal cost when data producers and data scientists are organizationally fragmented. New challenges derive from a common factor: the massive increase in data production capacity by the varied experimental devices, with impact on the perennial issue of data sharing, on the traditional practices for scientific credit, and on the modern, ontology based methods for the definition of knowledge.

We next consider three major issues encountered by scientists with respect to the management of data.

Data sharing and interoperability. The first problem is the difficulty to share data among scientists. The synergy in sciences requires a steady flow of communication between scientists [4]. In particular, it requires being able to reproduce results by following the same protocol in order to validate some new knowledge. But the exponential curve of digital technologies profoundly alters the data production capabilities of run-of-the-mill laboratories. The amount of data that can be produced is potentially enormous, possibly concerning thousands of variables. Such volumes add a further incentive to data sharing, beyond scientific goals: splitting the data processing costs (human, machines and software) between different labs. *So data sharing has become essential.*

However, important informations are in fact not sufficiently accessible.Because of the quantity of data, sharing has to rely on computers. Unfortunately, this may become impossible or to the least very complicated because of interoperability issues between the software of the different labs. The experimental instruments are often associated with proprietary commercial software, or poorly documented home-made software tools that export data in their own format.

The related focus for the LinkedWiki Platform is to build data sharing best practices on top of structurally non interoperable proprietary software.

Lab notes and scientific credit. A second problem appears very rapidly in the data production process when data are shared, the problem of scientific credit. The process of knowledge creation is accompanied and interwined with that of credit awarding, typically achieved via publications and patents. Scientists may be willing to share the raw data they produce, under various modes (embargo, limited access...), but they want to keep the benefit of their production effort as a competitive advantage for scientific discovery.

In the context of scientific credit, a tool is of particular importance, namely, the *lab-notebook* that records the history of findings. They can serve as a proof of discovery in case of controversy on anteriority. In this notebook, the scientist describes in details his/her research with, in particular, a total protection of the "intellectual (meta)data" about the data they produce. Integrating the lab-notebook with the data production process is critical for interoperability. Without such an integration, the scientists have to describe what they do in a paper notebook for the protection of intellectual property and in a computer system for describing the data production protocol, including experimental setup as well as further post processing. Besides the waste of precious time, this also result in inconsistencies, and is the cause of incompleteness of the paper, the computer version, or both. This jeopardize experimental reproducibility even with complete sharing of raw data and artefacts.

The integration of lab notes (including paper lab notes) in the production of scientific results is thus crucial for reproducibility. The LinkedWiki Platform helps the labs share safely their private lab notes as well as the corresponding data.

Discovery of data sets. The third and last issue is the discovery of datasets available to the community. The scientist and various software the scientist uses should both be able to discover relevant datasets. For this, we need to use standards for ontologies, and for ontology alignment.

This will be a main focus of the LinkedWiki platform that will rely on the use of modern standards in Web knowledge management.

2.2 Standards for data formats and knowledge

The variety of data formats reflects the history of scientific data. The obvious solution is standards, that we briefly describe in relation with description, search, and conceptual organization.

Description. The International Organization for Standardization, is working on developing standards for sharing workflow data (experimental and others) at a generic level. This organization has produced a first standard of particular importance, namely, DOI (Digital Object Identifier). As the name tells, DOI allows uniquely identifying objects. A DOI is associated to some metadata including a location, such as URL, where the object can be found. Currently, DOI are often used for sharing compressed files consisting of a publication together with a few artifacts.

The World Wide Web Consortium (W3C), a standardization organization, is proposing *Linked Data*, for publishing Web metadata. The LinkedWiki Platform relies heavily on this standard. Linked Data is based on the following concepts:

1. Conceptual things are unambiguously denoted by universal resource identifiers (URI).

2. URIs that are identifiable on the Web (starting with HTTP) denote objects that can be referred to and looked up ("dereferenced") by people and user agents.

3. When a URI is looked up, its "semantics" includes links to other related things (using their URIs). That semantics is represented in a standard format, namely Resource Description Framework (RDF).

The use of RDF enables that of the SPARQL query language. Furthermore, a new recommendation by the W3C, namely the Linked Data Platform describes a platform for sharing files including data with metadata with Linked data formats.

Search and discovery. DOI provides a mechanism for sharing generic objects and Linked Data exposes metadata, but they do not directly address search and discovery. Google provides Sitemaps and Schema.org technologies for indexing data via portals (e.g., publications in scientific journal).

This is rather limited, and it is rather hard to discover scientist workflow data of interest on the web only. Another means of discovering workflow data is via portals such as Datahub.io.With such portals, scientists can publish their open datasets and search for datasets in their domain. However, these portals are for open data only. In the context of a university or a lab that produce workflow data with restricted access, search for the discovery of workflow data of interest is typically not yet supported.

Ontologies. The LinkedWiki platform uses first the DCAT ontology (Data Catalog Vocabulary). As the European Union selected DCAT for its Open Data platform [10], it presents some guarantees of sustainability and can be expected to achieve a wide coverage. However, DCAT offers a minimal level of description, based on keywords only. Each dataset is also described using the RDF-based SPARQL 1.1 Service Description. This allows describing the means to access data using particular services. The service also refers to specific RDF ontologies that actually describe the data. For example, to publish multi-dimensional statistical data, the scientists may use the RDF Data Cube Vocabulary.

3. STATE OF THE ART

In this section, we mention other platforms that help scientists identify and locate online data repositories.

Catalogues. Many manually maintained catalogues are available on the Web that list repositories of scientific data, for instance the Open Access Directory [11] or the Awesome Public Datasets [14].

Some catalogues have the ambition to gather or automatically harvest large numbers of datasets. They typically provide search engines. One can cite for instance Re3data, DataCite, OpenDOAR and Openarchives. The standards facilitates the generation of these catalogues.

Open data platforms.

Generic platforms can also be used to manage data. There are principally five open-source data management platforms with instances running at research and government institutions, namely CKAN, DSpace, Figshare and Zenodo [2].

DSpace, Figshare and Zenodo have the principal aim to preserve and share the research outputs. With their search engines, it is possible to do a research in function of output's type like the type "dataset" but generally, the scientists use the to share their documents: publications, thesis, etc.

On the other side, CKAN has clearly the vocation to share all kinds of data, i.e. all origins [5]. It has the widest community of developers, thus offers more tools for managing data than other platforms and is exploited in institutional platforms e.g. Datahub.io or Data.bris. CKAN rely on the DCAT ontology and their metadata is exported in RDF format. However, the Web sites with CKAN's instances rarely deployed SPARQL endpoints. The use of their metadata by other platforms is therefore limited.

More importantly, it has been argued [2] that these platforms do not sufficiently cover the entire scientific workflow.

ODIPP and global approaches.

The DCAT ontology has been proposed by the European Commission for building the Open Data Interoperability Platform & Protocol [9, 10, 6]. The aim is to provide a homogeneised access to metadata descriptions of open datasets via a single point of access. For doing that, the European Commission encourages the publication of dataset's metadata as Linked Open Data via CKAN or other platform. Then, the ODIPP plateform can copy all these metadata and so provide a single access point via SPARQL.

Overall, the technologies exist for the management of scientific data and metadata. However, the present solutions are not fully satisfactory, and, in particular, are not well-adapted to the scientific workflows. The LinkedWiki platform is a contribution to the search for fully flexible solutions.

4. THE APPROACH

Attaching metadata to data sets. The platform encourages the publication of metadata about the scientific datasets. For each dataset, it is required to provide: the level of privacy (private, university open or open), the title of the dataset, a short textual description, a contact mail, links to the relevant branches of sciences, and links to concepts describing the dataset. All these are mandatory. A URL of a file or an API (with MIME type) may be provided. It is optional because one may want to reference some data before making them available on the Web.

The research information includes protocol, experimental, and artefact data, as well as corresponding documents (publications, courses, etc.). All these information together constitute a graph of information that is essential for the scientists. The links in this graph are described primarily in the protocol data. In general, these links are most useful to discover and reuse the data sets. Thus, the LinkedWiki platform helps the user build this graph in RDF, and then stores it. This graph forms an important part of the RDF metadata of the dataset. The RDF metadata (RDF graphs) are accessible (and can be queried) via a SPARQL endpoint.

The Linked Data approach allows attaching new metadata to some particular datasets without actually modifying or disclosing the original datasets. This is important (i) because one can enrich the dataset even without "owning" it, and (ii) even without altering it.

Giving a perennial URI to each dataset. To combine maximum flexibility with traceability, we associate a perennial URI to each dataset with the minimum of metadata that describe that dataset as well as its host and how it is hosted. So the scientists can freely choose how their data is hosted (even on Dropbox) but the platform can automatically report on its availability.

Selecting datasets with Wikipedia. One we have associated a perennial URI and metadata to a dataset, the next issue is facilitate the discovery of this URI. For this, the LinkedWiki Platform relies primarily on Wikipedia.

The publication of a dataset is typically too complex for the vast majority of scientists, who don't know the standards for publishing metadata and don't master the ontologies needed to qualify datasets. The LinkedWiki Platform offers a practical solution to publish data with the minimal ontological know-how, that is sufficient to facilitate their discovery by other scientists.

To see how, let us go the final goal (making data easily

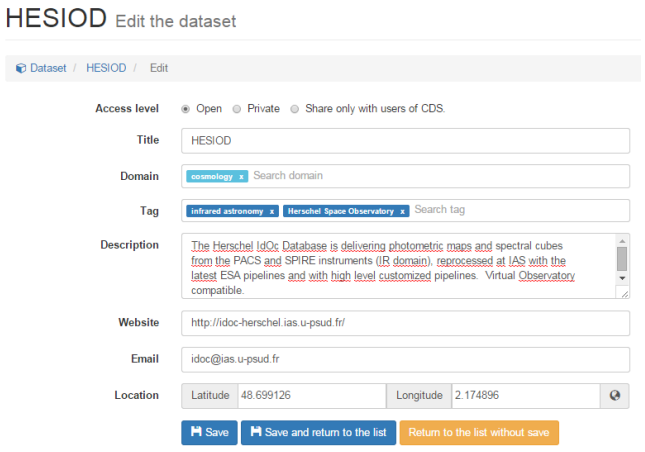

Figure 1: Form for attaching new metadata.

discovered) and observe how scientists search for datasets. Because of popularity ranking in web search engines, Wikipedia pages are very frequently in the top results even for very precise query terms and very specialized scientific topics [8]. It turns out that this Wikipedia page is often a reasonable description of the datasets they a scientist is looking for. If not, which is typically because the query terms were not accurate enough, the scientist is encouraged to rephrase the query until satisfied with the Wikipedia page.

So let us assume a scientist found the page of Wikipedia that corresponds best to the datasets of interest. We would like to minimize the time it then takes to actually find these datasets. With the LinkedWiki platform, they are *immediately* available. To the platform users, the available datasets are presented directly on the corresponding page in Wikipedia, thanks to a Wikipedia plugin we developed that customizes the appearance of Wikipedia articles. The links to relevant datasets is provided by a SPARQL query over the platform knowledge base. This approach allows easy discovery by scientists of datasets they ignore, while staying very close to their usual behaviors.

For this Wikipedia-based search for dataset to work, we have to ask the scientists to first relate their datasets to one or more Wikipedia pages. Wikipedia is available in 288 languages. For a majority of search engines in all languages, and most standard query terms, the corresponding scientific topics have a principal page on the Web and that page is a Wikipedia article. So by attaching a Wikipedia page to a dataset, the scientist attaches a unique *concept*, beyond any specific language. These are unique perennial URI within the project Wikidata (the knowledge base of Wikipedia). Thus, most standard research concepts now have a unique URI in Wikidata. This is an alternative to publishing the dataset on a platform with several specific classification systems such as JEL-Code [3] and ACM [1]. The classification system we use in LinkedWiki is much more general because it is built on Wikidata.

5. IMPLEMENTATION

This section describes the main software components we developed.

5.1 The publication APIs

Beyond a visual presentation of the metadata, the LinkedWiki platform offers a a SPARQL endpoint for directly querying them. Figure 1 illustrates the interface to input new metadata for a dataset. This published metadata can be easily reused thanks to the SPARQL endpoint, for instance by its use in Wikipedia discussed in Section 5.2.

As already mentioned we use the DCAT ontology. We extended DCAT with two new properties, *item* and *theme*. The *item* value is the URI of a Wikidata entity. It plays the same role as the *keyword* property in DCAT. The Wikipedia entity is simply more precise. We kept *keyword* for interoperability. The *theme* value is again the URI of a Wikidata entity that corresponds to a context of use. For example, a dataset about the *Herschel Space Observatory* satellite will have for theme the Wikidata entity for infrared astronomy. In summary, the *items* describe the "What?" and the *themes* describe the "What for?".

Considering the data formats and endpoints (termed *forms* in DCAT), LinkedWiki supports simple ones to make datasets available from different points of views.

To select scientific topics, it is not possible to use the standard API of Wikidata because a topic the user is looking for may be absent from Wikidata. LinkedWiki thus proposes its own Topic selection API. With it, a user can first select a topic that is already in Wikidata. However, a user who wants to select a topic that does not yet exist is offered the means to create a new page in Wikipedia and to enter the new concept in Wikidata. The concept then becomes visible in the API for selecting topics.

From an implementation viewpoint, the metadata are stored in an SQL database. An RDF materialized view and its SPARQL endpoint are supported. Presently, only RDF metadata for the open access level is exported. We intend to install a Virtuoso [12] database that will allow providing an entry point for the entire RDF graph with proper access control.

5.2 Direct access to datasets via Wikipedia

The direct access to datasets described in section 4 is implemented by some Javascript code that modifies the display of the Wikipedia page for those who installed the plug-in. The plug-in is easy to install and usable on all Wikimedia Foundation projects (Wikipedia in all languages, Wiktionary, etc.). We used standard Wikipedia technologies [13]. The code does the following: (i) find the Wikidata ID of the current page, (ii) generate a SPARQL query, (iii) display the datasets in the results (as clickable links).

The main goal of this development was to facilitate access to the datasets by scientists or even students by relying on a Web service they are familiar with, namely Wikipedia. This development also illustrates what can be done by combining Wikipedia and Linked data (here via the LinkedWiki Platform). We believe that this will encourage the lab developers to use Wikipedia to access their knowledge bases.

5.3 Extension of MediaWiki to query RDF data

When a scientist inputs information in a lab notebook, we would like to connect this information (enrich it) with available knowledge bases. This would lead to more modern lab notebooks, more appropriate to describe the data workflows, and in particular the artefact data. Another advantage of such an approch [7] is users contributions to improving the quality of knowledge bases. For labs notebooks, we en-

Figure 2: Use case where a wiki can show the last earthquakes.

courage the use of an extension of MediaWiki, one of the most popular open source Wiki. The extension, namely MediaWiki/LinkedWiki allows connecting the notes to the Wikidata ontology or to RDF graphs. With this extension, users can write a SPARQL query directly in the Wiki page. The query can be edited using a query editor. The query result is inserted in the page.

To realize the extension, we developed PHP SPARQL 1.1 client library. Figure 2 (from a thesis work) shows a MediaWiki where a researcher enriches a map using earthquake knowledge bases.

6. CONCLUSION

We described the LinkedWiki platform for scientific data sharing, that is based on standard Web technologies. We discuss how the platform helps scientists discover and reuse data.

From this first experience, we can draw some conclusions.

The first impact is the one that was most expected: the discovery of datasets by scientists (within the university) is greatly facilitated. They are able to discover these data automatically starting from the LinkedWiki SPARQL endpoint. In the long term, we hope to encourage massive data analysis (big data style) to acquire scientific knowledge but we are not yet there. In particular, raw data remains very hard to manipulate even for experts. The question of data interoperability remains posed.

Second, scientists could see the interest there is in making their datasets directly reachable from Wikipedia. They could get a first exposure to the semantic Web and to the Linked Data technology. They start understanding the interest of these new technologies, and are encouraged to use them more in the future.

A last impact still has to be verified. We believe that the use of ontologies and Wikidata by the scientists will lead to enriching the ontologies, building more links between them, reconciling them. This will eventually result in more interoperability between scientific data sources. This aspect is not yet clear to most scientists even, if it can be expected that they will be most affected by the outcomes.

We are intending to improve the platform in the following directions:

1. We will increase the amount of metadata that is automotically associated by the system. For instance, we will provide tools to analyse datasets and suggest metadata about them.
2. We also started helping scientists produce artefact data in interoperable formats.
3. We will provide a subscription system that will alert scientists of the reuse of their data by other scientists, and of their references in other data sets.
4. We will provide, in particular for teaching purposes, an environment for testing SPARQL queries on large RDF graphs, or on some topic-specific RDF datasets.

7. ACKNOWLEDGMENTS

This work has been partially funded by BorderCloud, the FUI project TIMCO, France Grilles and by the Paris-Saclay Center for Data Science (funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02).

8. REFERENCES

[1] ACM. Association for Computing Machinery classification system, 2012.

[2] 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*, pages 101–111. Springer, 2015.

[3] A. E. Association. JEL Classification System, 1969.

[4] B. A. Fischer and M. J. Zigmond. The essential nature of sharing in science. *Science and engineering ethics*, 16(4):783–799, 2010.

[5] T. O. K. Foundation. CKAN, the world's leading open-source data portal platform, 2015.

[6] S. Goedertier. Description of DCAT application profile for data portals in Europe. *Europa.eu*, 2013.

[7] M. Jaillard, S. Schicklin, A. Larue-Triolet, and J.-B. Veyrieras. A Comprehensive Microbial Knowledge Base to Support the Development of In-vitro Diagnostic Solutions in Infectious Diseases. In *I-SEMANTICS (Posters & Demos)*, pages 55–59, 2013.

[8] M. R. Laurent and T. J. Vickers. Seeking health information online: does Wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.

[9] D. C. of the European Commission. Repository of the Open Data Interoperability Platform, 2014.

[10] European Union. european union open data portal, 2015.

[11] P. Suber. The Open Access Directory, 2014.

[12] O. Virtuoso. *RDF Graphs Security with Virtuoso's database*, accessed April 3, 2015.

[13] Wikipedia. Global user pages. *Wikipedia, The Free Encyclopedia*, 2015.

[14] C. Xiaming and Many others. The Awesome Public Datasets, 2015.