Dendro: Collaborative Research Data Management Built on Linked Open Data

João Rocha da Silva[1]([✉]), João Aguiar Castro[1], Cristina Ribeiro[2],
and João Correia Lopes[2]

[1] Faculdade de Engenharia da Universidade do Porto/INESC TEC, Porto, Portugal
{joaorosilva,joaoaguiarcastro}@gmail.com
[2] DEI—Faculdade de Engenharia da Universidade do Porto/INESC TEC,
Porto, Portugal
{mcr,jlopes}@fe.up.pt

Abstract. Research datasets in the so-called "long-tail of science" are easily lost after their primary use. Support for preservation, if available, is hard to fit in the research agenda. Our previous work has provided evidence that dataset creators are motivated to spend time on data description, especially if this also facilitates data exchange within a group or a project. This activity should take place early in the data generation process, when it can be regarded as an actual part of data creation. We present the first prototype of the Dendro platform, designed to help researchers use concepts from domain-specific ontologies to collaboratively describe and share datasets within their groups. Unlike existing solutions, ontologies are used at the core of the data storage and querying layer, enabling users to establish meaningful domain-specific links between data, for any domain. The platform is currently being tested with research groups from the University of Porto.

1 Introduction

Research data is diverse and requires specific knowledge to be interpreted, driving user communities to create metadata recommendations. Metadata for datasets, as for any other kind of resource, requires a tradeoff between a comprehensive description and control of the production cost [8]. This is more drastic in the "long-tail of science" as institutions often lack financial resources for data curation [4]. As metadata schemas grow to encompass the needs of different groups, their descriptors may become unnecessary or irrelevant to others, even in similar domains, leading to an overall lack of interoperability [1,2]. This motivated some research groups to adapt and combine sets of descriptors from several metadata schemas in order to suit the needs of their applications, creating *Application Profiles* [3] to describe research datasets.

We focus on data description in the early stages of research, much like ADMIRAL [5], and propose that researchers choose their own set of metadata descriptors from existing ontologies. Dendro, our platform, innovates by integrating research datasets in the Semantic Web and allowing users to describe them

using concepts captured in ontologies. We combine this dynamic approach with
the advantages of a triple-based data model proposed in the same context [6]. To
simplify the workflow, we do not attempt to represent the contents of files as sets
of RDF triples (as done in VoID[1] for example) instead focusing on describing
and relating the files and folders themselves.

Dendro is designed to support researchers in their daily data management
activities. With a generic data model that allows on-demand metadata descriptor
selection by the user, it is completely built on both generic and domain-specific
ontologies. OpenLink Virtuoso and SPARQL are at the core of its data layer,
enabling metadata descriptions to be exposed on the Web and queried through
Virtuoso's SPARQL endpoint.

2   Enabling Collaboration and Interoperability

Dendro was designed from the start as an user-friendly interface layer for users
without data management knowledge. Users build a knowledge base using ontolo-
gies *in the background*, allowing them to focus on choosing the *properties* with
the right semantics for their descriptions without being concerned with design
and implementation issues that arise from ontology use. Given its collaborative
nature, the solution can be classified as a semantic wiki built on a triple store.
It differs from other semantic wikis like Semantic Mediawiki, for example, that
stores amalgamated sets of triples as "pages" in its relational database. Accord-
ing to the documentation[2], Semantic Mediawiki can use a triple store to provide
a SPARQL endpoint, but the synchronization between the relational database
and the triple store uses dedicated business logic—a trait shared by other linked
open data compatible systems.

Based on our own past developments in Semantic Mediawiki [7], we concluded
that its interface is not designed to allow users to combine descriptors from
several ontologies when describing a page[3]. Dendro, on the other hand, makes
it easier to describe any kind of resource using combinations of descriptors not
specified *a priori*. The ontology-based data model enables data management
personnel without coding skills to contribute by building and loading additional
ontologies into *their* Dendro, which can then be shared on the web to document
the descriptions and reused by others in the Dendro instances that they manage.

3   A Walkthrough of the Solution

In this section we will provide an overview of the main features provided by
Dendro in its current form. We demonstrate the usage of Dendro in the daily

---

[1] http://www.w3.org/TR/void/
[2] http://semantic-mediawiki.org/wiki/Help:Using_SPARQL_and_RDF_stores
[3] A *description template* must be specified *a priori* for each type of description page.

Fig. 1. Using Dendro to describe a mechanical engineering dataset

research data management activities within research groups from two very distinct domains—fracture mechanics experiments (mechanical engineering) and pollutant analysis (analytical chemistry)[4].

Figure 1 is a composite of screenshots showing how Dendro can be used to describe a dataset from the mechanical engineering domain. Area **1** shows the *project list* that allows users to see the projects that they have created in the system (i.e. there is an instance of `dcterms:creator` in the graph, with the project as its subject and the user as its object). Area **2** shows the main description interface. Note the list of options available to the user (area **2A**, from left to right: create folder, upload file(s), download folder, backup folder, restore folder, and show/hide deleted files). The file list **2B** shows the contents of the current folder and allows the user to navigate in the system. The *autocomplete* box **2C** is used to retrieve descriptors from the ontologies currently loaded in the Dendro instance, based on the values of their `rdfs:label` and `rdfs:comment` annotation properties—upon selection, the descriptor is added to the description area to be filled in. All descriptors originate from ontologies available on the web. Upon

---

Fig. 2. A free-text search and SPARQL query over Dendro's graph

loading an ontology into Dendro, its properties become available in the search box, provided they have their own `rdfs:label` and `rdfs:comment` annotation properties.

The system also provides a set of *smart* descriptors **3**, usually presented below **2C**, which can be seen as shortcuts for fast selection of most recently used descriptors. Upon first use, the system will simply recommend the most used descriptors in the system. When the user selects a descriptor, the system will give preference to descriptors from the same ontology. When the user selects another descriptor from a different ontology, the recommendation is broadened to the descriptors from the now two *active* ontologies. All changes to descriptor values are versioned, as can be seen in area **4**. Finally, the system supports recursive backup and restore of directory structures (including metadata) through `ZIP` files. Area **5** shows the contents of a complete backup of the current project— note the `metadata.json` file at the root, which contains all the metadata for all resources in the project's directory tree.

Figure 2 shows the resource described in Fig. 1 among the results of a full-text search for the term "fracture mechanics" over the Dendro system (**1**). The search is powered by an ElasticSearch index that indexes every resource in the graph by its literals and that is continuously updated. Area **2** shows a partial view of the results of a SPARQL query used to retrieve the metadata for the same resource—SPARQL queries such as this are used internally by Dendro to retrieve and modify data in the underlying OpenLink Virtuoso graph database.

4   Conclusions and Future Work

Dendro is a research data management platform designed to provide researchers with a collaborative environment for storing and describing their datasets. Ontologies are used as sources for properties, picked by researchers to describe their research data.

Dendro differs from other research data management platforms in its "all semantic web" approach. By employing a triple-based data model and OpenLink Virtuoso, each resource can have an arbitrary set of descriptors. As they interact with the system, Dendro users are actually building a Linked Open Data graph of interconnected research-related resources, while data access is performed internally via SPARQL all accross the platform.

Dendro development is informed by the requirements of a panel of researchers from the University of Porto, and preliminary tests have shown a good match between their data management needs and the services of the platform. We regard it as an effective practical application of semantic web technologies, as well as a catalyst for the creation of domain-specific lightweight ontologies.

References

1. Castro, J., Ribeiro, C., Rocha, J.: Designing an application profile using qualified dublin core: a case study with fracture mechanics datasets. In: Proceedings of the DC-2013 Conference, pp. 47–52 (2013)
2. Chan, L.: Metadata interoperability and standardization - a study of methodology Part I. D-Lib Mag. **12**, 1–34 (2006)
3. Heery, R., Patel, M.: Application profiles: mixing and matching metadata schemas. Ariadne Issue 25, September 2000. http://www.ariadne.ac.uk/issue25/app-profiles/
4. Heidorn, P.B.: Shedding light on the dark data in the long tail of science. Libr. Trends **57**(2), 280–299 (2008)
5. Hodson, S.: ADMIRAL: A Data Management Infrastructure for Research Activities in the Life sciences. University of Oxford, Technical report (2011)
6. Li, Y.-F., Kennedy, G., Ngoran, F., Wu, P.: An ontology-centric architecture for extensible scientific data management systems. Future Gener. Comput. Syst. **29**(2), 1–38 (2013)
7. Rocha, J., Barbosa, J., Gouveia, M., Ribeiro, C., Correia Lopes, J.: UPBox and DataNotes: a collaborative data management environment for the long tail of research data. In: iPres 2013 Conference Proceedings (2013)
8. Treloar, A., Wilkinson, R.: Rethinking metadata creation and management in a data-driven research world. In: 2008 IEEE Fourth International Conference on eScience, pp. 782–789, December 2008