Abstract

Metadata and ontology repositories are critical to ensure the discovery of existing vocabularies and the reuse of vocabularies and/or individual properties. However, these infrastructures should take into consideration the decision making process and criteria for the selection of a vocabulary of individual concepts or properties. Usage data in particular are important and can reassure on the maintenance of the vocabulary by a third party. This data is to some extent available through dedicated tools, such as semantic search engines. We illustrate the need for integrating usage data in the vocabulary infrastructures in order to support the reusability of vocabularies and therefore interoperability and data usability in science.

1. Introduction

Research is becoming more and more data centric (Hey, 2012). Researchers make use of many datasets which are available on the Web, including social data and open data sets, leading to the use of tools for manipulating large quantities of data (Big data). At the same time the development of open science and open access to scientific publications and datasets has led scientists to integrate more and more of their production to the Web infrastructure. In this context, the description of datasets with common metadata models and domain ontologies to represent knowledge are critical to the reuse manipulation, representation, and exchange of data and knowledge. They constitute key challenges of the Web based research infrastructure.

Despite the widespread use of standard vocabularies, most implementers have a need for tailoring them or adding properties. Instead of creating a new homemade property, they can reuse an existing metadata property.

While metadata and ontology repositories provide information on the structural context of metadata properties, i.e., the vocabulary they come from, it is difficult to identify their maintenance conditions and their actual usage context. In this paper, we propose a mechanism to provide this type of metadata in order to inform metadata implementers in their choice.

2. Reusing metadata properties

Despite the dissemination of standard vocabularies and the availability of metadata registries and ontology repositories, the reuse of vocabularies and individual metadata properties reuse is still not matching expectations. This decreases significantly data reusability. We show these challenges through two projects on Open data reuse and on the dissemination of scientific resources through scientific wikis.

2.1 Reusing vocabularies to describe Open data sets

In the scope of a study on Open Data (Martin et al., 2013), we have gathered statistics on metadata sets published on the Public Data EU catalogue which federates multiple catalogues in Europe in order to identify the level of openness of datasets as well as the variety of data providers. The metadata collected in May 2013 on 17,027 datasets used 236 properties. Indeed the JSON interface of the catalogue provides data from original catalogues. A lot of information is therefore expressed
through many distinct properties suggesting that the metadata properties were not well reused from existing catalogues.

### TABLE 1. Distinct values by license field

<table>
<thead>
<tr>
<th>Fields</th>
<th>Number of occurrences</th>
<th>Fields</th>
<th>Number of occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>licence</td>
<td>140</td>
<td>License_summary</td>
<td>55</td>
</tr>
<tr>
<td>License</td>
<td>15</td>
<td>License_title</td>
<td>26</td>
</tr>
<tr>
<td>license_url</td>
<td>5</td>
<td>License_uri</td>
<td>2</td>
</tr>
<tr>
<td>License_details</td>
<td>34</td>
<td>License_url</td>
<td>14</td>
</tr>
<tr>
<td>License_ID</td>
<td>26</td>
<td>mandate</td>
<td>29</td>
</tr>
</tbody>
</table>

Licensing information for instance can be found in 10 different metadata properties (TABLE 1). 79% of the metadata records have a meaningful value for at least 1 of these metadata properties. Nevertheless, the usefulness of a metadata property with a value in only 1 data source is limited. This illustrates the lack of reuse of metadata properties.

#### 2.2. Reusing vocabularies for a semantic Wiki

In the scope of the Wicri network of scientific semantic wikis (Ducloy et al., 2010), a set of properties had to be defined to create the semantic layer of wikis. In this network, a given topic (e.g., "metadata") can be studied on several wikis, displaying several points of view (e.g., "Computer science" or "Information science"), and/or in several contexts (e.g., France, Luxembourg, Europe, etc). Thus, on each wiki, a specific ontology is needed, dealing with the wiki’s thematic, in addition to a common ontology in order to guarantee coherence and semantic interoperability in the whole network.

Many properties can be reused from existing ontologies. For instance, "Has PC member" comes from "semanticweb.org" to link a conference with a person, PC member for this event. In the same way, many terms can be reused from existing vocabularies, for instance EuroVoc as global vocabulary. The decision to reuse a term from a vocabulary depends on its suitability but also on its source (Publications Office of the European Commission) and its widespread adoption.

Nevertheless, identifying such properties is not always easy. A consistent work of customization is required. As with many examples of reuse of ontologies, special situations require adaptations (usually extensions). For instance, "Has PC member" must be extended if a large number of events get diverse committees. The Wicri framework, and especially the network, introduces a new set of problems. For instance, on a given wiki, it could be necessary to use simultaneously MeSH and EuroVoc, two vocabularies using common concepts, but with different terms (and different relationships). Next step in this process, we now intend to use the semantic features of Wicri in association with bibliographic metadata, as learning set for data mining among large corpora: this raises again new issues, displaying a completely new “landscape” of using metadata and properties. Thus before we reuse metadata properties, we need to ensure that both their maintenance and their usage context are relevant to our case.

Both these cases illustrate the limited property reuse beyond standard vocabularies and the decision making process when deciding to reuse a metadata property. Our hypothesis is therefore that by following the decision making process of metadata property selection, it is possible to increase property reuse.

#### 3. Usage data to support the selection of vocabularies

Vocabularies usually include definitions of concepts and properties, their identifiers, and their relation with other concepts and properties. Knowledge Organization Systems may include scope notes as well as hierarchical or associative relations. The presentation of metadata models and ontologies including the objective and context for which they were conceived can usually be retrieved from the Web.

However, this does not inform on either vocabulary maintenance or actual usability. This is usually deduced from the source (e.g., the Library of Congress may be considered more reliable for
maintaining vocabulary than a university laboratory which created a vocabulary for a particular project). Who creates, maintains and uses a vocabulary is particularly critical in order to enable data linkage for Linked Open Data for instance, since data publishers have to rely on the fact that the meaning of a concept they are using remains the same over time and that its description remains accessible. The lack of versioning mechanism represents a major weakness of the current Web infrastructure to support data linkage (Van de Sompel et al., 2010).

The ability to relate to a community of metadata implementers who have implemented a property, or vocabulary is in this regard very important. The popularity of the resource is however not provided in metadata registries and ontology repositories such as the NSDL registry, the Dublin Core Metadata Registry, and the TONES ontology repository.

In order to assess an open source software project, production data (e.g., new releases), usage data (e.g., number of downloads), and communication data (e.g., mailing list activity) can be analyzed (e.g., Wynn, 2003; Ahmed et al., 2010).

Regarding metadata models, properties, vocabularies, and ontologies, production data are represented by potential versions which can to a certain extent be found in metadata registries. There is usually no bug report system or module added to the metadata model. Communication data can be retrieved from mailing lists activity for instance which have to be analyzed. Usage data relate to the type and number of implementers and to the type and number of collections which have implemented the properties and vocabularies.

Usage data on vocabularies can be found through semantic search engines for properties and concepts which have been published using one of the Semantic Web standards, including RDFa, RDF, microformats, Schema.org, and OpenGraph for instance. The Sindice search engine for instance suggests known properties from a term (FIG. 1). It displays the number of documents it has indexed which contain this property as well as the list of documents which can provide initial insight on the type of actors who use this property.

FIG 1. Occurences of the <foaf:mbox> property on Sindice

4. Conclusion: including usage data in the data infrastructure

Usage data can therefore be found through semantic search engines for instance, while structured information on the vocabularies can be retrieved from metadata registries and ontology repositories. Registries should include tools to automatically capture usage data and potentially general information on communication data, such as the liveliness of mailing lists that support the vocabularies. We suggest integrating these two infrastructures in order to support the decision making process of the metadata model, property, vocabulary, and ontology selection, thus facilitating reuse and improving the interoperability and accessibility of data.

References


http://publicdata.eu
http://owl.cs.manchester.ac.uk/repository/
http://sindice.com