Ontology Recommendation for the Data Publishers*

Antoine Zimmermann[1]

Digital Enterprise Research Institute
National University of Ireland, Galway, Ireland
firstname.lastname@deri.org

Abstract. We present a process for recommending Web ontologies that exemplifies quality, based on criteria that stimulate their reuse and spreading among data publishers: best practices, support by publishers and applications. The quality of ontologies can be assessed in a semi-automatic peer-review process.

1 Introduction

RDF data publishing on the Web has gathered momentum in the last few years, thanks to a general effort to link open data all over the Web. Yet, this trend is slowed down by the difficulty to find appropriate terms for the data to be described. Indeed, apart from a handful of well known ontologies, there are no readily available, easily findable vocabularies for most of the domains that would be good candidates for publishing data in a standard, linkable way.

The typical problems faced by would-be data publishers are: (1) ontologies defining the domain of interest do not exist; (2) they exist but are difficult to find because developed by small groups for experimentation, lacking advertisement; (3) they exist and can be found but they are of poor quality, not complying with standards or best practices; (4) they exist and can be found but there are too many, of mixed quality, and it is difficult to assess which ones are appropriate for a specific use case.

To address the first issue, user-friendly ontology editors have been developed Swoop[1], Protégé[2], etc. But this mostly requires that ontology and domain experts publish more and more data and terminologies. We assume that this will naturally happen when Linked Data and Semantic Web technologies will reach a critical mass which trigger a virtuous circle. We will not address this issue here.

The second item is somehow addressed by Semantic Web search engines. Several of them have been proposed, such as Swoogle [1], Sindice [2], SWSE [3],

* I would like to thank the Pedantic Web Group (http://www.pedantic-web.org/) for their useful discussions, and more particularly Stéphane Corlosquet, Richard Cyganiak, Renaud Delbru, Alexandre Passant and Axel Polleres. This work is partly funded by Science Foundation Ireland (SFI) project Lion-2 (SFI/08/CE/I1380).

[1] http://code.google.com/p/swoop/

[2] http://protege.stanford.edu/

FalconS [4], Watson [5], OntoSearch [6], Ontosearch 2 [7]. Besides, ontologies can be gathered together in repositories [8, 9] that provide additional functionalities for maintaining them. This can address to some extent the third and fourth items because voluntarily submitted ontologies are more likely to be considered by their authors as being of sufficient quality rather than ontologies randomly retrieved from the Web. Moreover, the implemented functionalities may help correcting possible errors and eventually would only display formally valid terminologies. To address the last issue, it has been proposed to add, *e.g.*, Web 2.0-like rating and voting functions to search engines and repositories (*e.g.*, Revyu [10]). However, reviewing ontologies needs advanced knowledge in fields that the ontology users may not have, and the ontology experts are not necessarily inclined to judge ontologies in the same way as social website assess conversations, products, etc.

Therefore, we believe that to guarantee access to the ontologies that are available, well supported and of quality, there is a need for promoting them more actively. In this paper, we argue in favour of having a committee of experts analyse Web vocabularies with respect to their suitability as reusable terminologies for data publishers. As a result of this analysis and evaluation—which can be partly automatised—the committee would advertise the ontology as a "quality vocabulary" and recommend it for describing information in the field applicable to such terminology.

To present this idea in more details, we first discuss the criteria that a Web terminology should fulfil to be labelled as "quality vocabulary" (Section 2). Then, we describe a possible approach to implement such an evaluation and recommendation framework (Section 3). Finally, we show how this integrates with ontology repositories (Section 4).

2 Criteria for a recommended Web vocabulary

Our objective in recommending Web vocabularies is focused on helping data publishers to find adequate terms for describing their data. For this reason our proposed initiative distinguishes itself from other ontology evaluation activities that focus more on engineering, design and logical issues. Moreover, we do not pretend to assess the quality of the modelling of the domain of interest, which could only be judged by a domain expert. Also, we encourage small, lightweight ontologies, which are easier to assess, reuse and scale. In this section we discuss possible criteria for quality vocabularies, recommended for reuse over the Web of Data. More precisely, data publishers would expect vocabularies that are: (1) justified by use cases; (2) easy to reuse and publish; (3) well interoperable with published Linked Data. We analyse these requirements to determine the criteria for quality vocabularies.

Justifying the existence of the vocabulary. As a primary requirement, a vocabulary should be accompanied by a statement about its utility. This includes a general description of the vocabulary and its scope as well as, more importantly, its related use cases. To avoid too much subjectivity in deciding

the relevance of a vocabulary in terms of usage, it can be required that a Web vocabulary proposed for recommendation should be supported by at least some data publishers. We consider this criteria a very important one and would not recommend a vocabulary, be it very well designed, if nobody considers using it. Usage should not be restricted to a unique dataset, not even to a big one by a major player in the field. At least two independent parties should be using the terms, or there should be strong evidence that the terms will be used by several distinct publishers in the near future. As an alternative proof of relevance, the vocabulary publisher could claim potential adoption by showing precise examples of possible usage. *E.g.*, a geo-location vocabulary can be proven to be useful if the authors show that translating existing geographic databases into linked data can be done in an easy and straightforward way to create and publish multiple datasets at a low cost. Finally, the utility of the vocabulary can be demonstrated if existing applications are usefully exploiting the associated data. This can justify the recommendation of a vocabulary since all data conforming to it will interoperate with those existing applications.

Ease of reuse and publication. Since one of the goal of recommending vocabularies is to increase interoperability by reducing the number of heterogeneous terminologies, it is important that the vocabulary be reusable as easily as possible. To achieve this, the content should be made understandable by non-ontology experts. Thus, one of the criteria is the presence of clear labels and textual descriptions for each term in the ontology. Moreover, granularity and complexity should be in line with the use cases. Highly expressive or too specific ontologies should be discouraged if they are not seriously justified. Finally, publication of data conforming to the proposed vocabulary should be made easier, *e.g.*, by providing tools that automatise (partly or totally) the publishing process. For instance, FOAF and SIOC exporters make the creation of online community RDF metadata fully automatic when integrated in a content management system.

Interoperability. To ensure better interoperability, several guidelines have to be followed. Obviously, vocabularies should be published in a standard format, namely RDF(S) and OWL. Additionally, since vocabularies are themselves part of the Linked Data, they should follow the best practices in the field [11]. This includes, *e.g.*, URI dereferencability, entity naming schemes, or authoritativeness of term definition. A term definition is considered "authoritative" if it describes an entity which is in the namespace of the document describing it. Most of these practices can be checked automatically, using, *e.g.*, RDF:Alerts[3]. Moreover, vocabularies are also used to reason about the data, so special care must be taken with this respect. It is desirable to enable interoperability of the vocabulary with both OWL tools and RDFS tools. On the one hand, an OWL ontology can be made more RDFS-friendly by defining all classes as both `owl:Class` and `rdfs:Class`. Similarly, properties defined as `owl:ObjectProperty`, `owl:DatatypeProperty` and `owl:AnnotationProperty` should also be defined as

[3] `http://swse.deri.org/RDFAlerts/`

`rdf:Property`. On the other hand, RDFS terminologies should declare each term as either of the aforementioned types, unless a strong justification comes from the use cases. For instance, the Dublin Core vocabulary does not specify the type of its properties in order to preserve flexibility. Also, an OWL ontology should be kept compatible with OWL DL as much as possible, and any exception should be justified. More generally, vocabularies should use the least expressive fragment of OWL that fulfils the desired requirements.

3 Implementing the recommendation process

This section shows how a framework for recommending Web vocabularies could be implemented in practice.

We notice that most ontologies and Web vocabularies, especially the most popular ones, are developed by academic researchers (FOAF, SIOC, Good Relations, Music Ontology, etc.) Thus, it would be possible to incite the ontology builders to publish their creation through our quality assessment process by establishing a regular submit/review/accept-reject process. An ontology or terminology for data publishing offers a solution to a problem or fulfil a certain need. It can thus be seen as a scientific contribution. We propose to have a call for vocabularies with a review process and editorial constraints.

First, an automatic tool will verify the compliance of the submitted vocabularies to the well established criteria mentioned in Section 2. If these are matched, then a peer-reviewing process will be undertaken by the committee, considering what has been discussed in the previous section. A submitted ontology must be accompanied by a description that will be published together with the ontology. The description—which can take the form of an article—must explains the purpose of the ontology—not only its domain but also its scope and granularity as well as possible or existing applications using it. It must show the utility and the need for such a vocabulary. This is partly proven by the fact that existing (independent) datasets are already using the terms or publishers have committed to use it in the near future.

The descriptions should be kept understandable by non-ontology specialists, and technical details can be given if, and only if, it contributes to showing the utility and interoperability of the vocabulary. As a result of this process, the ontology is either deemed not suitable for recommendation or recommended as a "quality vocabulary". Rejected ontologies can be improved and resubmitted later, eventually leading to better quality of the vocabularies. A centralised Web site would advertise these ontologies, provide documentation about them and allow searching and browsing them. Such a central place could be an existing ontology repository, as discussed in the next section.

4 Integrating recommendations in an ontology repository

The recommended ontologies should be easily searchable, browsable as well as matchable. These are common tasks made possible by ontology repositories.

Therefore, the recommendation process that we described previously may be integrated into a repository that could possibly include other non-recommended ontologies. However, the recommended ones should be emphasised and all operation should be applicable to the quality ontologies only. Moreover, non-recommended vocabularies could be marked by specific labels indicating that, although they are not recommended, they validate some of the criteria for quality.

5 Conclusion

In this paper, we argued that data publishers should be guided in their choice of vocabularies by actively recommending them ontologies that are considered of quality. The recommendation should be based on criteria that span from support by existing publishers and applications to the compliance to the best practices in this field. Evaluating those criteria could take the same form as an academic publication process, with a review phase. As a first step, we would like to launch a workshop on this topic which would create, we hope, an incentive for researchers to design and publish quality ontologies for the Web of Data.

References

1. Finin, T., Ding, Z., Pan, R., Joshi, A., Kolari, P., Java, A., Peng, Y.: Swoogle: Searching for Knowledge on the Semantic Web. In: Proc. of AAAI 2005, AAAI Press / The MIT Press (July 2005) 1682–1683
2. Oren, E., Delbru, R., Catasta, M., Cyganiak, R., Stenzhorn, H., Tummarello, G.: Sindice.com: a document-oriented lookup index for open linked data. International Journal of Metadata, Semantics and Ontologies **3**(1) (2008) 37–52
3. Harth, A., Hogan, A., Umbrich, J., Decker, S.: SWSE: Objects before documents! In: Proc. of the Billion Triple Semantic Web Challenge. (2008)
4. Cheng, G., Ge, W., Qu, Y.: FalconS: Searching and Browsing Entities on the Semantic Web. In: Proc. of WWW 2008, ACM Press (April 2008) 1101–1102
5. d'Aquin, M., Sabou, M., Dzbor, M., Baldassare, C., Gridinoc, L., Angeletou, S., Motta, E.: WATSON: A Gateway for the Semantic Web. In: Poster session of the European Semantic Web Conference, ESWC. (2007)
6. Zhang, Y., Vasconcelos, W., Sleeman, D.: OntoSearch: An Ontology Search Engine. In: Proc. of AI-2004. BCS Conference Series, Springer (2004)
7. Thomas, E., Pan, J.Z., Sleeman, D.: ONTOSEARCH2: Searching Ontologies Semantically. In: Proc. of OWLED 2007. Volume 258 of CEUR Workshop Proceedings., Sun SITE Central Europe (CEUR) (June 2007)
8. Pan, J., Cranefield, S., Carter, D.: A lightweight ontology repository. In: Proc. of AAMAS 2003, ACM Press (July 2003) 632–638
9. d'Aquin, M., Lewen, H.: Cupboard - A Place to Expose Your Ontologies to Applications and the Community. In: Proc. of ESWC 2009. Volume 5554., Springer (June 2009) 913–918
10. Heath, T., Motta, E.: Revyu: Linking reviews and ratings into the Web of Data. Journal of Web Semantics **6**(4) (2008) 266–273
11. Bizer, C., Cyganiak, R., Heath, T.: How to Publish Linked Data on the Web. web published (July 2007)