A Reuse-based Lightweight Method for Developing Linked Data Ontologies and Vocabularies

María Poveda-Villalón
Ontology Engineering Group. Departamento de Inteligencia Artificial.
Facultad de Informática, Universidad Politécnica de Madrid.
Campus de Montegancedo s/n.
28660 Boadilla del Monte. Madrid. Spain
mpoveda@fi.upm.es

Abstract. The uptake of Linked Data (LD) has promoted the proliferation of datasets and their associated ontologies for describing different domains. Particular LD development characteristics such as agility and web-based architecture necessitate the revision, adaption, and lightening of existing methodologies for ontology development. This thesis proposes a lightweight method for ontology development in an LD context which will be based in data-driven agile developments, existing resources to be reused, and the evaluation of the obtained products considering both classical ontological engineering principles and LD characteristics.

Keywords: ontology, vocabulary, methodology, linked data

1 Motivation and Research Questions

The Linked Data (LD) initiative enables the easy exposure, sharing, and connecting of data on the Web. Datasets in different domains are being increasingly published according to LD principles[1]. In order to realize the notion of LD, not only must the data be available in a standard format, but concepts and relationships among datasets must be defined by means of ontologies or vocabularies[2].

New vocabularies to model data to be exposed as Linked Data should be created and published when the existing and broadly used ontologies do not cover all the data intended for publication. Based on the guidelines for developing and publishing LD [5], LD practitioners should describe their data (a) reusing as many terms as possible from those existing in the vocabularies already published and (b) creating new terms, when available vocabularies do not model all the data that must be represented. During this apparently simple process several questions may arise for a data publisher. This PhD thesis proposal aims to develop a lightweight method to guide LD practitioners through the process of creating a vocabulary to represent their data. The ambition is to maintain the advantages, whilst offering solutions to cover the insufficiencies. The proposed method will be mainly based in reusing widely deployed vocabularies, describing data by means of answering the following questions:

[1] http://www.w3.org/DesignIssues/LinkedData.html

[2] At this moment, *there is no clear division between what is referred to as "vocabularies" and "ontologies"* (http://www.w3.org/standards/semanticweb/ontology). For this reason, we will use both terms indistinctly in this paper.

1. Where and how can vocabularies be found?
2. Which vocabularies or elements should be reused?
3. How much information should be reused?
4. How to reuse elements or vocabularies?
5. How to link elements or vocabularies?
6. How should terms be created according to LD and ontological principles?

2 State of the Art

The 1990s and early years of the new millennium have witnessed a growing interest in methodologies that support the creation of ontologies. All these approaches have facilitated major progress, transforming the art of building ontologies into an engineering activity. However, the existing methodologies should be reviewed and adapted to support ontology development and evolution in the Linked Data context. For example, the well-known traditional methodologies as METHONTOLOGY [3, 2], On-To-Knowledge [9] and DILIGENT [7], as well as the NeOn Methodology [10], propose time and resource consuming activities instead of simple and (semi)automatic processes as is desirable in an LD application development. Other approaches propose agile methodologies for ontology development, however, these are often unsuitable when working with Linked Data. The eXtreme Method [6] assumes that the requirements are set at the beginning and are unchanging, which is unrealistic while working with ontology evolution using LD. Other works as the XD Methodology [8] consider Ontology Design Pattern as the only kind of resource to be reused during the development, or do not consider ontological resource reuse as is the case of RapidOWL [1] instead of basing the development in reusing the terms already available in the LD cloud. In addition, none of the above mentioned methodologies consider particular requirements of ontologies and vocabularies that are going to be used in a LD environment. Within the literature on Linked Data [5], some high-level guidelines have been outlined to create vocabularies; however no concrete processes and detailed guidelines have been proposed to carry out such a development. Therefore, to the best of the author’s knowledge there are no methodological approaches that help ontology developers to build ontologies or vocabularies to be used in an LD context taking into account its particular characteristics, following a lightweight approach and providing detailed methodological guidelines for the proposed activities.

3 Proposed Approach

This PhD thesis proposal investigates how traditional and heavy methodologies for the development of ontologies and ontology networks could be lightened and adapted to an LD context by considering its particular requirements. A lightweight method for ontology development with a data-driven approach will be created including techniques and tools to carry out each of the proposed activities. Ontology evaluation techniques according to LD principles and architecture will also be developed in a pattern-based way in order to make their application highly automated and reusable.

As Fig. 1 illustrates, the proposed method consists of a workflow of activities based on the data intended for publication. With the aim of following a data-driven ontology/vocabulary development rather than the competency question development [4], used in the majority of existing methodologies, this workflow starts with a term extraction activity. The next step is to search for available vocabularies and ontologies already used in LD following the approach proposed in [5]. Subsequently, through the next steps, the available resources will be selected and integrated in order to produce a first model describing the data, for which, methodological guidelines will be provided. Finally, this model will be completed, in case it does not cover all the data to be represented, and evaluated before publishing and making the data available.

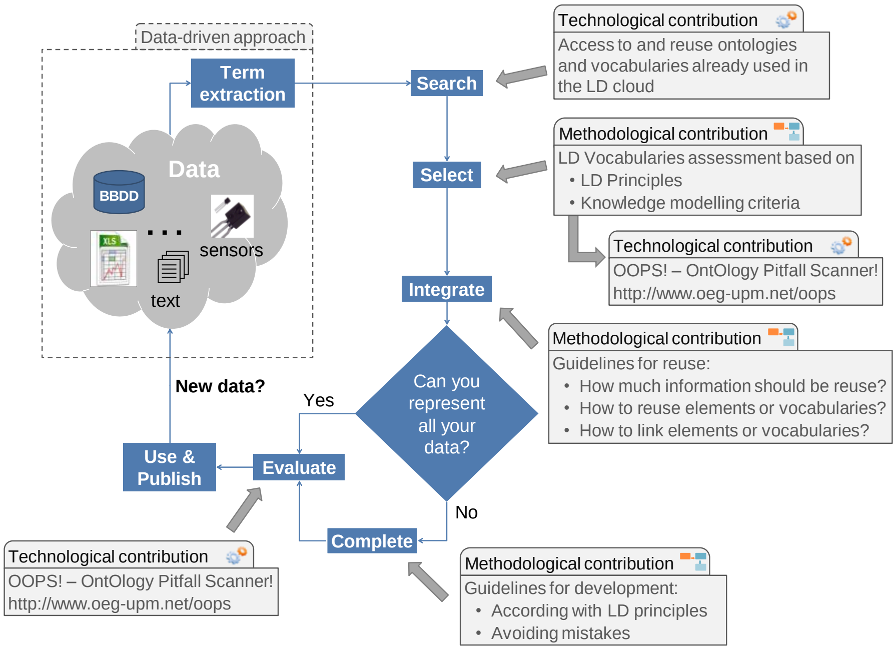
Data-driven approach
Term extraction
Data
BBDD
XLS
sensors
text
...
Search
Select
Integrate
Can you represent all your data?
Yes
No
New data?
Use & Publish
Evaluate
Complete
Technological contribution
Access to and reuse ontologies and vocabularies already used in the LD cloud
Methodological contribution
LD Vocabularies assessment based on
• LD Principles
• Knowledge modelling criteria
Technological contribution
OOPS! – OntOlogy Pitfall Scanner!
http://www.oeg-upm.net/oops
Methodological contribution
Guidelines for reuse:
• How much information should be reuse?
• How to reuse elements or vocabularies?
• How to link elements or vocabularies?
Technological contribution
OOPS! – OntOlogy Pitfall Scanner!
http://www.oeg-upm.net/oops
Methodological contribution
Guidelines for development:
• According with LD principles
• Avoiding mistakes

Fig. 1. Lightweight method for building Linked Data ontologies and vocabularies.

As well as the methodological contributions, Fig. 1 also shows the technological products that will be provided in this PhD thesis including technological support to access ontologies and vocabularies already used in LD cloud and OOPS!, an ontology validation tool. Table 1 presents the contributions that will be provided by this thesis in order to throw light upon the open questions presented in Section 1.

4 Planned Research Methodology

The research methodology to be followed will consist of prototype development from which a high level abstraction will be extracted. Results are evaluated at each iteration and used to inform the approach, which is fine-tuned until the results are satisfactory.

Initially, the state of the art in ontology development will be analyzed with a particular emphasis on problems involved with working in an LD environment.

Once the problem is defined, a first prototype of the agile method for ontology development presented in Section 3 will be proposed together with a first version of the technological application supporting it. This method will include prescriptive methodological guidelines for the proposed steps, namely, Search, Select, Integrate, Complete and Evaluate. These guidelines will be provided in a pattern-based manner whenever possible with the aim of enhancing their applicability and reusability.

Following this, an experimentation phase based on controlled experiments will be carried out over the obtained results taking as use cases the Ontology Engineering Group[3] LD developments. The experimental results will be used to improve the method and associated technological support. At least another iteration to evaluate the results and improvements will be carried out before proposing the final solution.

To conclude, the method will be analyzed a) from a user point of view by questionnaires to measure applicability; b) by controlled experiments involving the evaluation of ontologies developed for Linked Data applications carried out within the author's group (e.g: GeoLinkedData, UPMLinkedData, EcoGeoLinkedData, etc.) as initial evaluation environment and developments carried out by external organizations during a later evaluation step; and c) by comparison with other methods. The technological support will be compared with other tools with a similar purpose, gathering measures of: time spent by user in evaluating an ontology, usability tests, and user satisfaction after using the tool.

Table 1. Proposed steps, addressed open questions and PhD contributions.

Step(s)	Addressed Open Question(s)	PhD Contributions
Search	1. Where and how can vocabularies be found?	• Comparative study of the indexes or registries for ontologies used in LD cloud. • Techniques to access vocabularies.
Select	2. Which vocabularies or elements should be reused?	• Guidelines for assess and select vocabularies or elements to be reused. • Tool for evaluating vocabularies with respect to a set of modeling criteria
Integrate	3. How much information should be reused? 4. How to reuse elements or vocabularies? 5. How to link elements or vocabularies?	• Guidelines for ontology pruning and merging. • Guidelines for providing links between vocabularies and elements.
Complete Evaluate	6. How should terms be created according to LD and ontological principles?	• Guidelines for developing and enriching ontological terms according to LD criteria and ontological foundations. • Guidelines and technological support for ontology evaluation according to modeling criteria.

5 Conclusion

Describing data by means of vocabularies or ontologies is crucial for the Semantic Web and LD realization. LD development characteristics such as agility and web-

[3] http://www.oeg-upm.net/

based architecture force the revision and lightening of existing methodologies for ontology development. This paper briefly presents the motivation and the proposed approach of the thesis, the main goal of which is to propose a lightweight method for ontology development in an LD context following a data-driven approach. Such a method will be developed together with technological support to ease its application and will be based in agile developments and the evaluation of the obtained products considering both classical ontological engineering principles and LD characteristics.

The next steps consist of analyzing particular characteristics of LD developments and proposing a first prototype both for the method and its technological support. Following this, the obtained results will be evaluated in order to improve them in an iterative way.

Acknowledgments. This work has been supported by the Spanish project BabelData (TIN2010-17550). I would also like to thank Asunción Gómez-Pérez and Mari Carmen Suárez-Figueroa for their advice and dedication and Edward Beamer for his comments and revisions.

References

1. Auer, S.: *RapidOWL - an Agile Knowledge Engineering Methodology*. In: STICA 2006, Manchester, UK. 2006.
2. Fernández-López, M., Gómez-Pérez, A., Juristo, N. *METHONTOLOGY: From Ontological Art Towards Ontological Engineering.* 1997. Spring Symposium on Ontological Engineering of AAAI. Stanford University, California, pp 33–40.
3. Gómez-Pérez, A., Fernández-López, M., Corcho, O. *Ontological Engineering*. November 2003. Springer Verlag. *Advanced Information and Knowledge Processing* series. ISBN 1-85233-551-3.
4. Gruninger, M., Fox, M. S. *The role of competency questions in enterprise engineering*. In Proceedings of the IFIP WG5.7 Workshop on Benchmarking - Theory and Practice, Trondheim, Norway, 1994.
5. Heath, T., Bizer, C.: *Linked data: Evolving the Web into a global data space* (1^{st} edition). Morgan & Claypool. 2011.
6. Hristozova, M., Sterling, L. *An eXtreme Method for Developing Lightweight Ontologies*. CEUR Workshop Series, 2002.
7. Pinto, H.S., Tempich, C., Staab, S. *DILIGENT: Towards a fine-grained methodology for DIstributed, Loosely-controlled and evolvInG Engineering of oNTologie*s. In Ramón López de Mantaras and Lorenza Saitta, Proceedings of the ECAI 2004, August 22nd - 27th, pp. 393--397. IOS Press, Valencia, Spain, August 2004. ISBN: 1-58603-452-9. ISSN: 0922-6389.
8. Presutti, V., Daga, E., Gangemi ,A., Blomqvist E. *eXtreme Design with Content Ontology Design Patterns*. WOP 2009.
9. Staab, S., Schnurr, H.P., Studer, R., Sure, Y. *Knowledge Processes and Ontologies*. IEEE Intelligent Systems 16(1):26–34. (2001).
10. Suárez-Figueroa, M.C. Doctoral Thesis: *NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse*. Spain. Universidad Politécnica de Madrid. June 2010.