

Preventing Ontology Interoperability Problems Instead of Solving Them

Editors: Krzysztof Janowicz, Pennsylvania State University, USA and Pascal Hitzler, Wright State University, USA'

Solicited reviews: Giancarlo Guizzardi, Federal University of Espirito Santo (UFES), Brasil and Martin Raubal, University of California, Santa Barbara, USA.

Eero Hyvönen

Semantic Computing Research Group (SeCo), Aalto University and University of Helsinki, P.O. Box 15500, FI-00076 Aalto, Finland

Abstract. A major source of interoperability problems on the Semantic Web are the different vocabularies used in metadata descriptions. This paper argues that instead of *solving* interoperability problems we should focus more effort on *avoiding* the problems in the first place, in the spirit of Albert Einstein's quote "*Intellectuals solve problems, geniuses prevent them*". For this purpose, coordinated collaborative development of open source vocabularies and centralized publication of them as public vocabulary services are proposed. Methods, guidelines, and tools to facilitate this have been developed on a national level in the Finnish FinnONTO initiative, and are now in pilot use with applications and promising first results

Keywords: interoperability, vocabularies, ontology libraries, ontology services

1. Interoperability of Vocabularies

Much of the power of the Web comes from the freedom for anybody to publish and link his/her own content as the Web of Pages. When moving into the era of the Semantic Web, the Web of (Linked) Data, content is being linked on the level of ontological concepts and metadata underlying the pages¹ [3]. This leads to interoperability problems, especially interoperability regarding metadata schemas and vocabularies used for filling element values in the schemas. Approaches to schema interoperability include the dumb-down principle, as suggested in the Dublin Core (DC) community², and using a shared schema ontology onto which other metadata representations can be transformed, as suggested by the CIDOC CRM and FRBR communities³. In contrast, this paper focuses on interoperability problems

due to *domain vocabularies* (ontologies of hierarchically organized domain-specific concepts) used in annotations, not to schema models such as DC of CIDOC CRM that are also sometimes called "vocabularies" or "ontologies".

Content aggregated in semantic portals, or on the web scale in the Linked Data initiative, comes from actors and organizations that produce content for their own purposes and come from different disciplines, cultures, and countries. As a result, lots of different, partly overlapping vocabularies are used in metadata descriptions. To approach the interoperability problems, various techniques of ontology matching (mapping) [5] are used. For example, lots of mappings based on the *owl:sameAs* relation have been created for the resources in the Linked Data cloud. There are, for example, mappings between the place resources of DBPedia⁴ and GeoNames⁵. A key problem here is how to deal with situations, where

¹ <http://linkeddata.org/>

² <http://dublincore.org/>

³ <http://cidoc.ics.forth.gr/>

⁴ <http://dbpedia.org/>

⁵ <http://geonames.org/>

multiple entity names and identifiers are used for a single real world object [4], and where different objects have the same name or identifier. The same problem is encountered in Web 2.0 sites, where tagging using literals without identified meaning is causing more and more semantic confusion as more and more tags are being created (e.g. "jaguar" as an animal, or a car or an airplane model).

2. Coordinated Collaboration for Vocabulary Creation

The mess of meaning references on the metadata level on the Semantic Web creates lots of interesting research problems to study. Most research on interoperability issues seems to be focusing on developing methods and tools for obtaining interoperability between heterogeneous annotations (e.g. the datasets of the Linked Data initiative). However, from a non-academic practical viewpoint, this is a problem that should be avoided in the first place as far as possible. Obviously, more research effort should be focused on developing methods, tools, and practices by which metadata could be produced on a larger scale in an interoperable way at the time of creating it. Instead of solving interoperability problems we should rather try to prevent them by better ontology services, coordination, and collaboration in ontology development and content creation.

FinnONTO⁶ 2003–2012 is a research project and a Living Laboratory experiment [6, 7], where the idea is to establish a collaboration framework for vocabulary development and services on a national level for the Semantic Web. The main goal of FinnONTO is to create an open source, national level cross-domain "content infrastructure" for the Semantic Web, aligned with international vocabularies, standards, and practices. This infrastructure and network of concepts can be paralleled, on a conceptual level, with the construction of railroad, electrical, or telephone networks in the past.

The work is based on the domain independent Semantic Web standards⁷ of the W3C, such as RDF, SKOS, OWL, and SPARQL, but the heart of the system is domain-specific ontologies. While standardization work at W3C focuses on defining general principles of ontological structuring and reasoning, such as subsumption and inheritance, the

general goal of FinnONTO is to facilitate cross-domain interoperability of metadata descriptions on a domain-specific vocabulary level. The idea is that when content is published on the web, it should be possible to connect it semantically with other related (cross-domain) contents based on a system of mutually aligned domain ontologies.

The vocabulary infrastructure has been built by transforming nationally used traditional keyword thesauri [1] into lightweight ontologies, which makes the ontologies interoperable with already indexed content in databases. A key goal in the work is to encourage collaboration between ontology developers of different domains by proving a general FinnONTO ontology framework in which new vocabularies can be aligned with existing ones already during the ontologization process, instead of afterwards. The kernel of the FinnONTO system [7] is the General Finnish Ontology YSO developed from the widely used General Finnish Thesaurus YSA that consists of some 25,000 general concepts and that is maintained by the National Library of Finland. The corresponding ontology YSO has been extended by various domain-specific daughter ontologies, based on other national thesauri used in domains such cultural heritage, agriculture and forestry, applied arts, geography, photography, and others. These ontologies create together virtually one ontology, the Collaborative Holistic Ontology KOKO, that now has over 70,000 general concepts, not including ontology-like datasets, such as places, persons, mammal and bird species of the world, and historical events.⁸

Figure 1 illustrates the structure of KOKO, with the top ontology YSO on top, and overlapping domain ontologies AFO (agriculture and forestry), MAO (cultural heritage), TAO (applied art), and VALO (photography) extending its concept hierarchies.

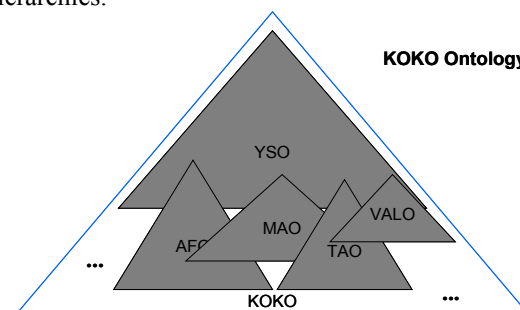


Fig. 1. KOKO system of overlapping aligned ontologies [6].

⁶ <http://www.seco.tkk.fi/projects/finnonto/>

⁷ <http://www.w3.org/standards/semanticweb/>

⁸ <http://www.seco.tkk.fi/ontologies/>

3. Commandments for Social Vocabulary Development

The key idea in the ontologization process is to aim at a system of vocabularies that are *born interoperable* with each other. To facilitate this, a new thesaurus is first matched with the general YSO top ontology in order to identify potential overlaps. The result is a Protégé editor⁹ project that includes YSO and the new thesaurus concepts. This structure is then corrected and maintained manually. (Alignment with other overlapping ontologies is also possible in a similar way.) In this way, the work already done in YSO can be reused in daughter ontologies and, at the same time, interoperability is enhanced by collaboration.

Vocabulary work in our view is as much a social process as it is a technical challenge. The work is guided by the following principles or "ten commandments":

1. Add machine semantics to legacy vocabularies. Start transforming thesauri [1] into machine interpretable lightweight ontologies in order to boost their usage on the Semantic Web.
2. Think cross-domain. Consider not only your own micro world, but also cross-domain usage of concepts when making ontological decisions.
3. Establish collaboration networks of domain expert groups. Nobody masters the whole universe.
4. Reuse others' work.
5. Maintain interoperability with the past and other ontologies. Otherwise benefits of collaboration are lost.
6. Proceed in small steps. Adding even little semantics can be very useful (and keeps e.g. the funding agencies happy).
7. Respect different ontological views. It is not possible to come up with only one ontological view of the world.
8. Accept imperfect models. The ontology will never be fully perfect.
9. Minimal ontological commitment. Keep ontological structures simple and generic in order to facilitate cross-domain reuse.
10. Coordinate the work and add new commandments if needed. This is done now by the FinnONTO research project but later,

⁹ <http://protege.stanford.edu/>

if the project is successful, by another coordinating organization.

4. Vocabulary Services for Legacy Systems

Another key component of the FinnONTO infrastructure is the National Ontology Service ONKI¹⁰ [15] hosting currently over 80 ontologies and vocabularies. The idea is provide the vocabularies as a free open source service for both human *and* machine users to utilize. ONKI ontology services such as concept finding, browsing, fetching, and query expansion [13, 14] can be integrated with legacy systems through REST, Web Service, or AJAX APIs in a way that is analogous to using Google Maps as an external service in applications. We hope that by making vocabulary services available and usable in an economically motivating way, organizations and people start using shared ONKI vocabularies and URIs, preventing interoperability problems rising from using local or depreciated vocabularies, and ambiguous literal terms in annotations. Other ontology servers on the web with the goal of publishing and sharing ontologies in public include Cupboard [2] and BioPortal [11].

5. Evaluation

The feasibility of the FinnONTO approach is tested and demonstrated in practice by applications, such as the collaborative semantic portals Museum-Finland¹¹, HealthFinland¹² [12], and CultureSampo¹³ [8] that makes use of the whole KOKO system aligned with some international vocabularies, such as the Getty vocabularies¹⁴ AAT, TGN, and ULAN. In summer 2009, 150 organizations in Finland and abroad had been registered to use ONKI services, and new ontologized vocabularies in the system have been developed by external organizations, e.g. an ontology for maritime terms (MERO) and for literature content (KAUNO). The latter one that has been used, based on the ONKI services, for annotating over 50,000 pieces of Finnish novels and short stories in a Web 2.0 fashion by Finnish

¹⁰ <http://www.onki.fi/>

¹¹ <http://www.museosuomi.fi/>

¹² <http://www.terveysuomi.fi/>

¹³ <http://www.kulttuurisampo.fi/>

¹⁴ http://www.getty.edu/research/conducting_research/vocabularies/

librarians for the semantic literature portal Kirjasampo¹⁵. In HealthFinland metadata is being created using the ONKI ontologies and services by a variety of national health organizations, and the system is in use¹⁶ and maintained by the National Institute for Health and Welfare since 2009 [12].

Our own experience suggests that gaining semantic interoperability in terms of vocabularies is a very tedious task and hinders fast publication cycle from legacy databases to the Web. In CultureSampo, for example, the content is harvested from tens of museums, libraries, archives, media companies, and web sources producing heterogeneous content. The vocabulary interoperability problem should in our mind definitely be addressed seriously at the time and place of content creation, rather than after harvesting the content, and we hope that the FinnONTO infrastructure is a step towards facilitating this in practice.

6. Discussion

Changing the established practices of vocabulary development, and adapting software in legacy systems to use ontologies cannot happen instantly but only over time. However, we believe there is now a promising road ahead to go based on the collaborative FinnONTO approach, although many problems of interoperable ontology development need to be addressed in the future.

A concern is the management of changes in the evolving ontologies and their alignments. Ontology versioning is needed because 1) the underlying real world or 2) our conceptualization about it may change [10], or 3) the underlying vocabulary standards evolve. Here one faces the problem that old content has been annotated using an old vocabulary while the end-user or applications may use a modern vocabulary or different old vocabularies. To address the problem, alignments between vocabulary versions along the temporal dimension are needed. An approach to modeling temporal ontology changes was developed in the Finnish Spatio-temporal Ontology SAPO¹⁷ modeling over 1000 geographical changes of Finnish counties (e.g. boundaries and names) since 1865 [9].

An important question in sharing ontologies is application specificity or point of view dependency.

¹⁵ <http://www.kirjasampo.fi/>

¹⁶ <http://www.terveysuomi.fi/>

¹⁷ <http://www.seco.tkk.fi/ontologies/sapo/>

An ontology developed from one point of view may not be usable from another perspective. To pursue application independence, the FinnONTO vocabularies are kept lightweight with as little ontological commitment to applications as possible. The vocabularies provide only little more than the skeletal RDFS subsumption hierarchy of concepts, and it is left up to the applications to build more domain specific semantics based on that.

End-users, domain experts, and ontology engineers may have different views to a domain. In such cases, different separate ontologies for the same domain may be needed, aligned with each other. For example, in the HealthFinland portal, the content is annotated using domain expert vocabularies, such as Medical Subject Headings (MeSH)¹⁸, but the vocabularies provided for the citizen end-users in the faceted search engine are based on layman's concepts extracted using a card-sorting technique [12].

Still another concern is whether two vocabularies sharing the same concepts should share the same ontological structure, too. Since there can be different views and opinions to modeling the real world, the modeling choices in a vocabulary in FinnONTO can be made independently from those in other overlapping vocabularies. The FinnONTO framework only makes the different vocabularies and views visible to all parties, encouraging but not forcing to sharing structures.

It is our hope that supporting collaboration in distributed ontology development facilitates cost-efficient creation of large cross-domain vocabularies with better interoperability than using a centralized approach or distributed development without coordination. It is also our hope that supporting social collaboration will lead to ontologies of better quality. By using the ONKI service for ontology publishing, the results of the joint efforts can be utilized in practical applications easily as ready-to-use services—both by human and machine end-users.

Acknowledgements

Thanks to Martin Raubal, Giancarlo Guizzardi, and Krzysztof Janowicz for fruitful comments concerning an earlier version of this article. Tens of researchers have been working in the various parts of the FinnONTO 2003–2012 project. The research has been funded by the Finnish Funding Agency for

¹⁸ <http://www.nlm.nih.gov/mesh/>

Technology and Innovation (Tekes), and a consortium of over 40 companies and public organizations.

References

- [1] Jean Aitchison, Alan Gilchrist, David Bawden: *Thesaurus Construction and Use: A Practical Manual*. Routledge, 2000.
- [2] M. d'Aquin, H. Lewen: *Cupboard—A Place to Expose Your Ontologies to Applications and the Community*. In: *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, Springer-Verlag, 2009.
- [3] C. Bizer, T. Heath, T. Berners-Lee: *Linked Data—The Story So Far*. In: *Journal of Semantic Web and Information Systems*, Vol. 5, No. 3, 2009.
- [4] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, Antonio Mana: *Entity Name System: The Backbone of an Open and Scalable Web of Data*. In: *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2008)*. IEEE Computer Society, 2008, pp. 554–561.
- [5] J. Euzenat, P. Shvaiko: *Ontology Matching*. Springer-Verlag, 2007.
- [6] Eero Hyvönen: *Developing and Using a National Cross-Domain Semantic Web Infrastructure*. In P. Sheu et al. (eds), *Semantic Computing*. IEEE Wiley, 2010.
- [7] Eero Hyvönen, Kim Viljanen, Jouni Tuominen, Katri Seppälä: *Building a National Semantic Web Ontology and Ontology Service Infrastructure—The FinnONTO Approach*. *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, Springer-Verlag, 2008.
- [8] Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuitinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkilä, Panu Paakkari, Joonas Laitio, Katariina Nyberg: *CultureSampo—Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user*. In: J. Trant, D. Bearman (eds): *Museums and the Web 2009: Proceedings*. Archives & Museum Informatics, Toronto.
- [9] Tomi Kauppinen, Eero Hyvönen: *Modeling and Reasoning about Changes in Ontology Time Series*. In: Rajiv Kishore, Ram Ramesh, Raj Sharman (eds): *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*. Springer-Verlag, 2007, pp. 319–338.
- [10] M. Klein: *Change Management for Distributed Ontologies*. PhD thesis, Free University of Amsterdam, The Netherlands, 2004.
- [11] M. Musen, N. Shah, N. Noy, B. Dai, M. Dorf, N. Griffith, J. Buntrock, C. Jonquet, M. Montegut, D. Rubin: *BioPortal: Ontologies and Data Resources with the Click of a Mouse*. In: *Proceedings of the Annual Symposium Proceedings/AMIA Symposium*, 2008.
- [12] Osma Suominen, Eero Hyvönen, Kim Viljanen, Eija Hukka: *HealthFinland—a National Semantic Publishing Network and Portal for Health Information*. *Journal of Web Semantics*, vol. 7, no. 4, 2009, pp. 271–376.
- [13] Jouni Tuominen, Matias Frosterus, Kim Viljanen, Eero Hyvönen: *ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services*. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer-Verlag, 2009.
- [14] Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, Eero Hyvönen: *Ontology-based Query Expansion Widget for Information Retrieval*. *Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009)*. *CEUR Workshop Proceedings*, Vol. 449, <http://ceur-ws.org/>, 2009.
- [15] Kim Viljanen, Jouni Tuominen, Eero Hyvönen: *Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI*. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer-Verlag, 2009.