

Policies for Public Domain Ontologies for the Intelligence Community

Elisa F. Kendall¹, Jay Jacobs², Deborah L. McGuinness³, and Stephen Schwab²

¹Sandpiper Software, Los Altos, CA (ekendall@sandsoft.com)

²SPARTA, El Segundo, CA (jay.jacobs@sparta.com, stephen.schwab@sparta.com)

³Stanford University, Stanford, CA; affiliation as of date of conference: Rensselaer Polytechnic Institute, Troy, NY (dlm@cs.rpi.edu)

Abstract

Numerous RDF vocabularies and OWL, KIF, and other knowledge representation language ontologies have been contributed to the growing body of ontologies available in the public domain over the last ten years. Many of these were created with government-funded research support in the US and EU. Only a small subset is reusable, and fewer are appropriate for use in applications supporting evolving Intelligence Community requirements. This is partly due to decreasing funding available in the US in particular, but also because of lack of well-specified policies for vocabulary management, metadata, and provenance specification. In this paper we will highlight some of the challenges we have faced in developing and attempting to reuse ontologies in support of DARPA and US Department of Defense initiatives, and provide fodder for discussion of requirements for public domain ontologies.

Introduction

Numerous RDF (Resource Description Framework [1]) vocabularies and OWL (Web Ontology Language [2]), KIF (Knowledge Interchange Format [3]), and other knowledge representation language ontologies have been contributed to the growing body of ontologies available in the public domain over the last ten years. Many of these were created with government-funded research support in the US and EU. Only a small subset is reusable, and fewer are appropriate for use in applications supporting evolving Intelligence Community (IC) requirements. This is partly due to decreasing funding available in the US in particular, but also because of lack of well-specified policies for vocabulary management, metadata, and provenance specification.

Many of the ontologies available from the Protégé library [4], the National Center for Biological Ontology [5], via Semantic Web Central [6], and other collections are domain-specific, focused, for example, on use cases in pharmacogenomics, radiology, or other biomedical or other domain-specific applications. Of those that are more general in nature and potentially relevant for intelligence use, many are incomplete due to funding limitations, reflect varying coverage and granularity, and/or were developed with very specific application requirements in mind. They rarely include the level of metadata and provenance necessary to meet IC requirements [7-8]. Even fewer provide sufficient metadata from a vocabulary management perspective to enable users to understand the ramifications of long-term dependence [9].

Our insights in requirements and methodology for ontology and vocabulary development and management for intelligence use are derived from experience on a number of DARPA, ARDA, other US Department of Defense and NOAA programs as well as commercial projects. They reflect discussions with colleagues in Object Management Group (OMG), World Wide Web Consortium (W3C), and related international standards activities as well as direct conversations

with and surveys of intelligence analysts. And, while individual researchers may have varying opinions on specific aspects of ontology development methodology, choice of language, tooling, and so forth, we have found little to no disagreement on critical issues in vocabulary management or metadata and provenance requirements.

Motivation

A number of the better known, publicly available RDF vocabularies and ontologies, including the OWL language itself and general metadata schemes such as Dublin Core [10] and the Simple Knowledge Organization System (SKOS)[11], were initially created by small teams of developers in collaboration with much larger user communities. It is possible that their utility is responsible for their popularity, but we believe this is also due to the commitment made by the developers to support their users, resulting in continuous improvement over time. In contrast, while the majority of the ontologies developed under the DARPA DAML program are the direct result of significant initial effort on the part of the research community, many of these are showing signs of age and reflect the limited funding available for specific ontology development even over the course of that program. For example, a number of projects, including the time zone ontology components [12] developed for use with DAML Time [13], OWL-S [14], and other domain-specific applications depend on the ontology components for ISO 3166 (codes for the representation of names of countries) available in the DARPA DAML library [15]. This particular ontology provides the set of the alpha-2 codes specified in ISO 3166-1 as of its publication date (2003), but has not been revised since and does not support a number of other data values present in the current standard, such as alpha-3 and numeric codes, references to administrative languages, and so forth. This information was likely not needed when the ontology was initially developed, and some of the detail has been added in a recent revision of the standard. The example highlights issues such as maintaining currency, documenting maintenance policies, describing development requirements, the authority of the publisher with respect to the original standard, and so forth, however, which are clearly important to those who might want to reuse these ontologies in other applications, and particularly for IC applications that clearly must be able to count on currency in this and many other “general” vocabulary subject areas.

Vocabulary Management

The Semantic Web Deployment Working Group has continued work initiated by the Semantic Web Best Practices and Deployment Working Group to publish some basic principles for managing RDF vocabularies and OWL ontologies based on experience with Dublin Core, SKOS, and other ontology development. Some of the most basic issues under discussion include:

- Naming conventions, including use of URIs and publishing ownership and commitments to URI persistence
- Documentation – for example, following the strategies used for Dublin Core, SKOS, and others
- Maintenance policies
- Version management strategies
- Publishing the formal schema (in addition to the documentation)

These represent only the tip of the iceberg, however, in consideration of requirements for utility in IC applications in our view. For certain ontologies, such as those reflecting ISO standards that are published and managed by a formal registration authority, such as the Library of Congress for ISO 639 (language codes) and ISO 3166, we believe that ontology publication should become the responsibility of the registration authority. It is much more likely that members of the IC would trust an ontology published by the registration authority for the standard, or other publicly recognized authority for a particular subject matter (NIST, for example, with regard to units of measure and related standards), than most other potential publishers such as a small company.

Ontology-based applications for operational IC use also require significant metadata reflecting definition provenance, currency, accuracy, completeness, and a development process that is closer to software engineering CMMI-level 3+ compliance than a typical research program would entail.

We believe that from a practical perspective, development of policies for ontology and vocabulary development and management must be established prior to considering development of such public domain resources.

References

- [1] Dan Brickley and R. V. Guha. RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation, World Wide Web Consortium, Amsterdam, the Netherlands, 10 February 2004. Latest version is available at <http://www.w3.org/TR/rdf-schema/>.
- [2] Mike Dean and Guus Schreiber, eds., Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL Web Ontology Language 1.0 Reference, W3C Recommendation, World Wide Web Consortium, Amsterdam, the Netherlands, 10 February 2004.
Latest version is available at <http://www.w3.org/TR/owl-ref/>.
- [3] M. R. Genesereth & R. E. Fikes, Knowledge Interchange Format, Version 3.0 Reference Manual. KSL Report KSL-92-86, Knowledge Systems Laboratory, Stanford University, June 1992.
- [4] Protégé Ontologies Library, 2007.
<http://protege.cim3.net/cgi-bin/wiki.pl?ProtegeOntologiesLibrary>.
- [5] National Center for Biomedical Ontology, BioPortal Ontology Library, 2007.
<http://www.bioontology.org/>.
- [6] Semantic Web Central, 2007, <http://www.semwebcentral.org/>.
- [7] Pinheiro da Silva, P, McGuinness, D., McCool, R. 2003. Knowledge Provenance Infrastructure. IEEE Data Engineering Bulletin 26(4), pp. 26-32.
- [8] Christopher Welty, J. William Murdock, Paulo Pinheiro da Silva, Deborah L. McGuinness, David Ferrucci, Richard Fikes. Tracking Information Extraction from Intelligence Documents. In *Proceedings of the 2005 International Conference on Intelligence Analysis (IA 2005)*, McLean, VA, USA, 2-6 May, 2005.

- [9] Kendall et al., eds. Basic Principles for Managing an RDF Vocabulary, W3C Semantic Web Deployment Working Group draft, 2007.
<http://www.w3.org/2006/07/SWD/wiki/VocabMgtDraft>.
- [10] The Dublin Core Metadata Initiative (DCMI) and DCMI Metadata Terms, 2007.
See <http://dublincore.org/>.
- [11] Simple Knowledge Organisation System (SKOS), 2007.
See <http://www.w3.org/2004/02/skos/>.
- [12] Feng Pan and Jerry R. Hobbs. A Time Zone Resource in OWL.
<http://www.isi.edu/~pan/timezonehomepage.html>.
- [13] Hobbs et al., DAML Ontology of Time. <http://www.cs.rochester.edu/~ferguson/daml/>.
- [14] Martin et al., OWL-S 1.2 Pre-Release. <http://www.ai.sri.com/daml/services/owl-s/1.2/>.
- [15] Dean et al., DAML Country Codes. <http://www.daml.org/2001/09/countries/>.