Ontology Evaluation - " Beauty in the eye of the beholder ? "

Christopher J. O. Baker[1], Robert H. Warren[2], and Volker Haarslev[1]

[1] Concordia University, Montreal, Quebec, Canada, [2] University of Waterloo, Waterloo, Ontario, Canada

Introduction

A renewed focus on ontologies results from the recently agreed standard format for formal specifications of knowledge, namely the Ontology Web Language (OWL) [OWL www.W3C.org]. The promise of being able to exploit the richness of conceptualisations for a multitude of computational applications has resulted in an increase in the creation of ontologies and a new generation of computational ontologists. This community is concerned with the correctness, usability and content of ontologies.

Categories of researchers interested in evaluating ontologies include: philosophers, ontology developers, computational linguists, biologists and logicians and their criteria are varied. Much debate now focuses on the criteria by which ontologies should be evaluated and this is often driven by the usefulness of ontologies for a given application. Applications of ontologies are increasingly cited in the following categories:

- Ontology to support natural language processing / text mining
- Ontology to support annotation of genes for gene expression analysis
- Ontology to facilitate data integration
- Ontology as cornerstones of the semantic web
- Ontology for inference / knowledge discovery
- Ontology as an educational tool / resource

Criteria for ontology evaluation can be divided into domain dependent criteria and domain independent criteria. This poster highlights these evaluation criteria / metrics and provides case study evaluations of ontologies available over the Internet.

Philosophical Evaluation

An ontologist is concerned with:
The philosophical correctness of the conceptualisation of knowledge in ontology.

A recent evaluation of axioms within the Gene Ontology identified frequent violations of rules key to accurate knowledge representation [Smith, Koehler and Kumar 2004]. These are described below:

Univocity – Terms should have the same meanings (and thus point to the same referents) on every occasion of use. During ontology development, the same entity is often assigned different names. e.g., in the Gene Ontology the *part-of* relation represents three different meanings.

Positivity – Complements of classes are not themselves classes. The absence of a property is sometimes considered a distinction, though this is not an ontological class in itself. e.g., *non membrane-bound organelle* may be a class but *not a membrane-bound organelle* cannot.

Objectivity – The classes that exist are not a function of the current state of our biological knowledge. Terms such as 'unclassified' or 'unknown ligand' do not designate natural biological concepts. If we have no knowledge about a concept is it a concept?

Single Inheritance – No class in a classificatory hierarchy should have more than one parent on the immediate higher level.

Exhaustiveness – The classes on any given level should exhaust the domain of the classificatory hierarchy. This means that our representation of knowledge should exhaustively cover the whole domain.

Intelligibility – The terms used in a definition should be simpler (more intelligible), more logical or ontologically basic than the term to be defined.

References

Smith, B., Köhler, J. & Kumar, A. (2004). On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. E. Rahm (ed.), *Database Integration in the Life Sciences* (DILS 2004). Berlin: Springer.

López-Pérez, A. (2004). Integración de la aplicación OntoMetric en la plataforma WebODE. Proyecto fin de carrera. Facultad de Informática, Universidad Politécnica de Madrid.

Hartmann, J., Spyns, P., Giboin, A., Maynard, D., Cuel, R., Suárez-Figueroa, M. C. & Sure, Y. (2005). Methods for ontology evaluation. *EU-IST Network of Excellence (NoE) IST-2004-507482 KWEB.*

Brewster, C., Alani, H., Dasmahapatra, S. & Wilks, Y. (2004). Data Driven Ontology Evaluation. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2004)*, Lisbon, Portugal.

Maedche, A. & Staab, S. (2002). Measuring Similarity between Ontologies. In *Proceedings of the European Conference on Knowledge Acquisition and Management – EKAW-2002.* Madrid, Spain. LNCS/LNAI 2473, Springer, pp. 251-263.
Guarino, N. & Welty, C. (2002). Evaluating Ontological Decisions with OntoClean. Communications of the ACM: Vol. 45, No. 2.

Haarslev, V., Moeller, R. & Wessel, M. (2004). Querying the Semantic Web with Racer + nRQL. In *Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL'04)*. Ulm, Germany.

Domain Dependent Evaluation

Domain Content Evaluation

Ontologies lie at the heart of the semantic web, a vision that encompasses the utilisation of a multitude of ontologies available across the World Wide Web. Integration of domain conceptualisations at common semantic junctions where domain content can be matched is a critical step. Domain content is the primary criterion when considering an ontology for a specific application, when interrogating the knowledge within the conceptualisation or for evaluating an ontology for reuse in a new knowledge engineering project. Though tools like Ontometric (which computes quantitative evaluations for the latter goal based on 160 characteristics describing the ontology domain [Lopez-Perez 2004]) do exist, a thorough evaluation of domain coverage and domain content in ontologies across the web is currently lacking.

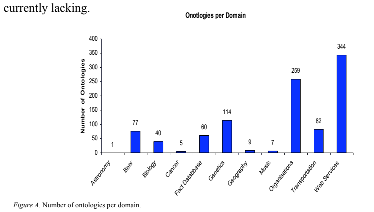


Figure A. Number of ontologies per domain.

Given the existence of multitudes of ontologies across the internet we sought to evaluate domain content and features of ontologies accessible with computational query tools. A domain content evaluation of 1600 ontologies was made by parsing OWL files for URI tags, depositing them in a relational database and querying them with simple keyword searches. A rudimentary semi-manual classification of ontologies based on their domain content was made. Most frequently occurring were ontologies representing web services, organisations such as governments and corporations, followed by genetics and transportation ontologies (Figure A). Within the 'identifiable' categories, the number of entities (concepts, roles and instances) were counted. This identified genetics as the domain most richly represented (Figure B).

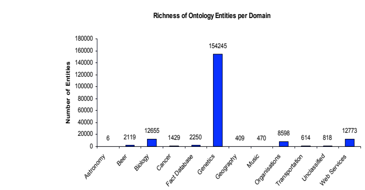


Figure B. Number of ontology entities per domain.

Domain Application Evaluation

Given their rich conceptualisations with advanced semantics, ontologies can provide crucial contributions to specific applications such as Natural Language Processing and text mining. NLP tools can populate ontologies with instances extracted from texts; the efficiency of such tools are considered in evaluating the ontology. Additionally, the compatibility of an ontology's hierarchically-controlled vocabulary to the corpus under examination are also evaluated. Metrics to measure these properties are summarised below and in [Hartmann et al. 2005].

Precision – (the number of correctly instantiated items as a percentage of all those instantiated) and **recall** (correctly instantiated items as a percentage of the total number of correct items in a 'gold standard' instantiated ontology) are used in ontology instantiation studies.

Cost based evaluation metric – Error rates instead of precision and recall are used [Hartmann etal 2005] in ontology instantiation studies.

Vector space models – of instances in a corpus and in ontology are used to measure the fit between ontology and text corpus terms [Brewster etal 2004].

Lexical Comparison Level measure – [Maedche and Staab 2002] compares the contents of two ontologies without considering their conceptual structure.

Domain Independent Evaluation

OWL Construct Evaluation

The Ontology Web Language provides three increasingly expressive sub-languages, which contain language constructs, some of which are unique to the sub-language. We parsed 1,600 OWL files and identified the frequency of occurance of each of the constructs in each of the ontologies. Figure C shows the percentage of OWL ontologies that used the particular language construct. OWL LITE constructs are shown in blue while OWL DL/FULL constructs are shown in red. Typically those constucts of OWL LITE employed exist with approximately double the frequency of OWL DL/FULL constructs. RDFS tags were not considered.

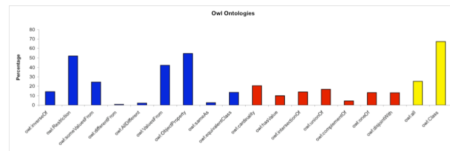
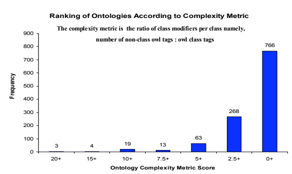


Figure C. Percentage of OWL ontologies per particular language construct

Occurrences of OWL constructs in each of the 1,600 ontologies were further counted and the ratio of class features (concepts) to properties and modifier features (roles and attributes) was computed to suggest a metric representing complexity of the ontology. Ontologies were ranked according to this metric. Figure D shows the number of ontologies per category of the metric score. Most ontologies scored between 0 and 2.5 implying that the majority of ontologies available across the internet are of low complexity. Only 7 ontologies scored above 15. Many of these ontologies originate from organisations with many years of ontology development experience. Although further validation of the metric is required, this may be a useful exercise in addressing what OWL features should be considered in an evaluation metric.



212186.owl DARPA / BBN Technologies, Metric Score 13
216166.owl Drexel University, Metric Score 14
13435.owl Fujitsu, Metric Score 17
213939.owl Mindswap, University of Maryland, Metric Score 18
215689.owl eBiquity, University of Maryland, Metric Score 22
216804.owl Mindswap, University of Maryland, Metric Score 24

Figure D. Ranking of ontologies according to complexity metric score.

Formal Evaluation

OWL files are representations of formal ontologies and their taxonomical structures can be compared with a predefined ideal taxonomical structure to detect inconsistencies. This is the methodology behind OntoClean [Guarino and Welty 2002] and relies on the ontological notions of rigidity, unity, identity and dependence. OntoClean evaluates whether constraints specified in an ontology are violated within the ontology. These notions are below:

Rigidity – A specification of essentiality. *A property is rigid if and only if it is necessarily essential to all its instances.*

Unity – A description of the way parts of an object are bound together, so that it is clear what is part of the object and what is not. *An individual is a whole if and only if it is made by a set of parts unified by a relation.*

Identity – The ability to identify individual entities as being the same or different. Identity criteria are conditions used to determine equality (sufficient conditions) and are entailed by equality (necessary conditions).

Dependence – Can an entity exist alone (independent), does its existence imply the existence of something else (rigid dependence), does it imply the existence of some entities that are instances of a specific class (generic dependence) or does a property holding for X depend on something else besides X (property dependence)?

OntoClean's formal evaluations provide useful insight into semantic models, yet these insights are structural and formally driven and do not allow us to infer anything about the usability of the ontology.

Domain Independent Evaluation

Description Logic Evaluation

By taking advantage of the DL usable formalism of OWL-DL, ontology evaluations can be conducted using description logic based queries posed in a query language like nRQL using a reasoning engine like RACER [Haarslev, Moeller and Wessel 2004]. Given the formal structure of the ontology, elementary, yet highly relevant queries can interrogate the ontologies to reveal: the classification hierarchies of ontologies, depth in ontologies, numbers of concepts, roles, instances, average number of child concepts and multiple inheritance. These metrics can indicate the level of complexity within ontologies suggesting their maturity and their suitability to support knowledge discovery through reasoning and inference. Table 1 shows the results of posing such queries to 50 randomly selected OWL ontologies.

OWL Filename	Depth of Ontology	# of Multiple Inheritance	# of Child ren per Concept	# of Concepts	# of Roles	# of Individuals	Complex ity Metric
212186	3	0	0.13	8	29	0	13.38
212192	3	0	0.46	13	34	1	10.89
212200	3	0	0.47	17	40	1	10.74
212766	6	3	0.72	61	48	13	1.84
212816	3	0	0.44	9	40	0	1.00
212886	3	0	0.47	17	40	1	10.74
212998	3	0	0.25	16	18	4	0.08
213072	4	0	0.46	13	40	0	2.94
213435	4	1	0.59	17	100	0	17.00
213507	2	0	0.00	3	26	0	11.00
213519	3	0	0.07	27	70	0	10.13
213939	2	0	0.00	2	56	10	18.00
214501	4	1	0.43	19	40	0	3.90
214759	2	0	0.00	1	20	3	3.00
214879	3	0	0.47	17	40	1	10.74
214881	3	0	0.47	17	40	1	10.74
214907	3	0	0.43	7	24	6	3.10
214909	3	0	0.17	6	28	5	5.75
215288	3	0	0.47	17	40	1	10.74
215462	3	0	0.42	12	60	0	4.40
215528	4	0	0.33	40	114	4	3.95
215627	3	0	0.25	4	78	0	10.25
215689	2	0	0.00	4	54	2	22.00
215719	6	4	1.00	24	99	7	3.70
215796	3	9	0.32	223	556	2	1.38
216042	4	0	0.80	15	20	2	0.03
216128	4	0	0.38	13	62	0	9.04
216136	4	0	0.54	13	34	0	5.57
216154	9	7	0.89	109	216	52	5.00
216166	3	0	0.21	14	58	0	14.36
216170	4	0	0.33	142	647	551	8.55
216174	6	0	0.75	100	230	41	4.68
216511	3	0	0.57	7	36	4	10.33
216557	4	0	0.50	8	24	6	6.09
216561	4	0	0.67	39	360	0	13.91
216587	3	0	0.17	6	28	5	6.50
216758	3	0	0.08	12	167	11	10.72
216772	3	0	0.34	29	74	0	4.23
216780	4	0	0.22	23	88	5	8.13
216804	3	0	0.33	3	50	0	24.00

Table 1. DL computed metrics from OWL ontology files

Conclusion

Ontology evaluations are gaining importance in light of efforts to employ ontologies in a multitude of application scenarios.

Given the number of opinions and perspectives, evaluation criteria are a *mixed bag*. It is too early to be able to point to any single metric or criteria that will assist as a generic ontology evaluation tool.

Challenges lie ahead in translating evaluation criteria into usable metrics.

Acknowledgments

Genome Quebec funds the authors of this work through FungalWeb: "Ontologies, the Semantic Web and Intelligent Systems for Genomics"