Ontologies: Where are we at?

Robert Warren[1,*]
rhwarren@uwaterloo.ca

Christopher O. J. Baker[2]
baker@encs.concordia.ca

1 School of Computer Science, University of Waterloo, Waterloo, On, Canada
2 Department of Computer Science, Concordia University, Montreal, PQ, Canada
* Work done while visiting Fungal Web Project

Abstract

Since the announcement of the creation of a Web Ontology Working Group [1] in November of 2001, two markup standards, DAML+OIL and OWL, were introduced in December of 2001 and March of 2002 respectively. We were concerned with the adoption rate of on-line ontologies by organisations.

We knew of few projects that were using ontologies to support work beyond trivial examples and wanted to get evidence of actual use of ontologies by organisations. Using data from Google [2], the Internet Wayback Archive [3] and custom crawlers we were are able to observe a rapid increase in ontology usage in the last 2 years.

The adoption pattern appears to be a power curve, in a manner similar to most Internet technologies. We review the granularity of the available ontologies as well as the distribution within each of the potential application fields. A simple classification method is used to provide an understanding of the fields currently making use of large ontolog ies. We highlight contribution made by ontologies of biological domains.

Introduction

Since the success of the GO [4] consortium and systems biology initiatives in reversing the reductionism bias in the study of biology the development of ontologies has begun to be recognised as a valuable pursuit (e.g.: Snomed [5], GO, OBO [6]). Furthermore, the introduction of the OWL Standard from the w3c in 2001 and press coverage / scientific publications citing this ontology standard have become commonplace.

Concurrently, search tools designed specifically to find semanticly enhanced tools and or files have become available [7]. Much of this work has gone unnoticed by the biology community; hence the initial questions asked by a biologist are about the prevalence of ontologies and what they can be used for. Because of the relative youth and its distributed nature, these questions are difficult to answer.

Given current research, we can however partially answer questions about their use. Ontologies have been used for instantiating domain knowledge, as a resource for advanced query generation and as mediating documents for data exchange. The ontological standards provide a flexible mechanism to represent knowledge in a machine readable format that explicitly documents assumptions, constraints, equivalences and organisation.

Establishing the availability of an appropriate ontology for a specific domain and coverage is also difficult. Beyond the canonical 'toy' ontologies, not only will individual users evaluate the same ontology differently for the same application, but the different stake-holders (ontology developer, end user or philosopher) will have different evaluation requirements.

Finally, an interested person might ask what criterias should be used to evaluate an ontology in the context of the global semantic web. Much of the promise in modern ontological research lies in the possibility of integrating knowledge across ontologies.

These are all questions with complex answers and we outline some initial insights into ontology publication, development and quality. We did this by tracking a number of ontologies available on the Internet over time and analysing their contents. The specific questions that we wished answered were:

- How many ontologies are out there? How large are they?
- What are the author's objectives in writing an ontology? What domain were they targeting?
- How integrated are the available ontologies? What is the potential for a semantic integration engine?

Methodology

We set out to create a data set of publically available ontologies and study their contents and evolution over time. Our objective was to determine what the current state of ontology usage 'in-the-wild' was and what work was being done beyond trivial, 'toy' examples.

The ontologies were found using both daily Google queries and a free-running web crawler that searched over 2 million web pages. The ontologies were retrieved and stored with time-stamped data. In each case, we attempted to query the Internet Way-back Archive for previously archived copies of the ontologies. When available, this data gave us two interesting pieces of information: the earliest point in time where the ontology was made available and any changes that occurred to the ontologies.

Because of the data sources that are used in this study, some bias is inserted into the analysis. We do not feel that the bias is critical, but we document it for completeness.

It is to be expected that many more ontologies are available either commercially or internally in organisations. However, we concern ourselves only with publically advertised ontologies. For these reasons, our dataset may not be a completely representative sample of ontologies in actual use.

Furthermore, we used the Internet Way-back Archive for historical data about ontologies. Because the Archive updates its database on a six-month cycle, the activity in the six months before August 2005 may be under-reported. Of course, the Archive does not have complete coverage of the web at all time periods; this may also dampen the reported availability of ontologies.

Organisations using ontologies

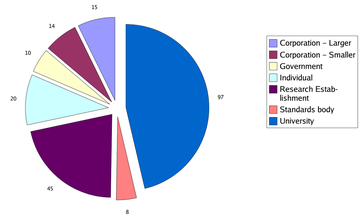

Figure 1 – Number of organisations within each classification.

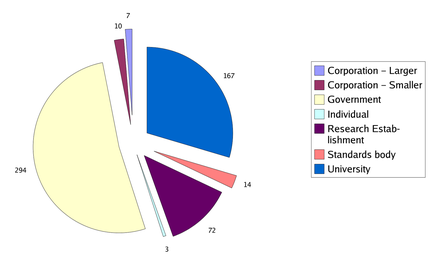

Figure 2 – Amount of information published by each type of organisation.

We classified the organisations publishing the ontologies into several categories that we felt were convenient to represent the current actors. The categories were Corporations (Large and Small), Governmental Institution, Individuals, Standards Bodies, Universities and non-academic Research Institutions.

Here we broke down the number of institutions within each category which were publishing ontologies. Unsurprisingly, Universities and Research Establishments make up the majority of organisations which are publishing ontologies on the web. We though it interesting to add the 'Individuals' category in that a number of ontologies are being published by people who are interested in the topic personally. They are also the third largest publishing group.

Alternatively, if we use the same grouping to look at the absolute amount of data published, Governmental organisations coming in at first rank. This is primarily due to several heath related ontologies released by US Government institutions. These are closely followed by Universities and Research Establishments.

Ontologies available on the web

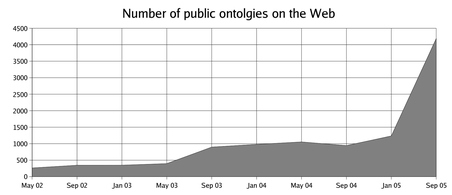

Figure 3 – Number of observed ontologies available.

Figure 3 plots the number of ontologies observed to be available over time on the world wide web. The first official release of the Daml+oil ontology standard was in December 2001, while the Owl standard was released in Feb 2004.

Interestingly, the rise in the number of ontologies available is similar to a power curse; a behaviour that is consistent with the adoption of most Internet technologies.

The question then becomes whether the growth will be sustained and whether the ontologies created are small, 'me too' example, ontologies or a generalised adoption of the technology.

Ontology size and complexity

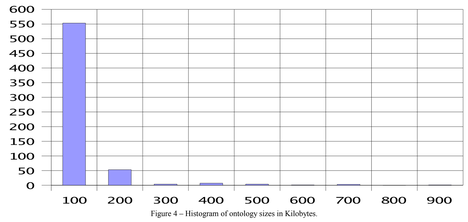

Figure 4 – Histogram of ontology sizes in Kilobytes.

A concern that we had was with the size and complexity of the available ontologies. We were concerned that only small examples would be available, with no evidence of complex problems being solved using the technology.

Figure 4 is a histogram of the ontology sizes in Kilobytes. As is to be expected, the great majority of the ontologies were small in size, with the number of ontologies inversely proportional to their size.

However, at the tail end of the histogram a few very large ontologies are available, such as Go, Fungal Web [8] and NCI Cancer [9], that are over 15 Megabytes in length. As all three ontologies are released by different organisations, we at least see the beginnings of their use for moderately complex systems.

Ontology creation and maintenance

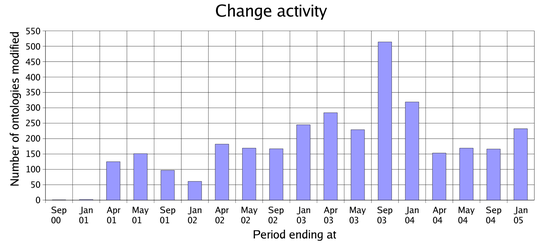

Figure 5 – Number of ontologies observed to have been changed within the time period.

We previously looked at the availability and the size of the ontologies, but we also have change information on the ontologies available on the website. It is to be expected that a number of these ontologies are on-going projects that require modifications and additions.

Hence, Figure 5 is a plot of the number of ontologies which changed during the time period. The figure must be interpreted carefully as a small bias exists in the graph because of the creation of new ontologies. Similarly, the decrease in activity in the last 6 months can be explained by the Wayback Archive policy of not publishing data before a 6 month maturity.

With this in mind, an analysis of the graph reveals that ontological development is very active. It is not possible to determine if this is a reflection of their use in a production setting or in on-going experimentation. However, it is an indication that not all the available ontologies are stale, static remnants of a technological fad.

References

[1] The World Wide Web Consortium (W3C) http://www.w3.org/
[2] Google seach engine http://www.google.com/
[3] The Wayback Archive http://www.archive.org/
[4] The Gene Ontology Consortium http://www.geneontology.org/
[5] Snomed Clinical Terms http://www.snomed.org/
[6] Open Biomedical Ontologies http://obo.sourceforge.net/
[7] J. Mayfield and T. Finin. (2003) Information retrieval on the Semantic Web: Integrating inference and retrieval, SIGIR Workshop on the Semantic Web.
[8] A. Shaban-Nejad, C. Baker, G. Butler and V. Haarslev (2004) The FungalWeb Ontology: Semantic Web Application for Fungal Genomics, 1st Canadian Semantic Web Interest Group Meeting (SWIG04).
[9] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, B. Parsia, and J. Oberthaler. (2003) The national cancer institute's thesaurus and ontology. Journal of Web Semantics, 1(1).
[10] V. Haarslev, R. Moeller and M. Wessel. (2004) Querying the Semantic Web with Racer + nRQL. Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL'04).

Query Tool

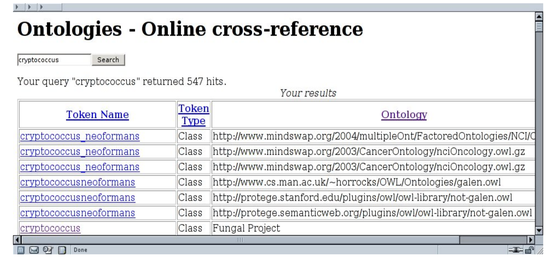

Figure 6 – Screen shot of web query tool for ontology object names,

As a means of exploring the contents of the available ontologies, we developed a simple query interface over the web (Fig. 6) that allowed us to search for strings within the ontologies. In this way, we were able to query all ontologies for similar concepts, irrespective of the structure of the ontologies.

The database of all queried entities in the ontologies found on the Internet has made it possible to search for commonality between ontologies. Using this tool we are able to identify potential semantic junctions between ontologies. Table 1 provides a brief overview of what information is currently available within the database for a set of selected biology-related terms.

Term	Number of entities containing term.	Number of ontologies containing term
Enzyme	249	16
Fungus	21	1
Cryptococcus	547	6
Membrane	472	8
Blood	339	10
DNA	733	23

Table 1 – Occurance counts of terms in entities and ontologies.

We did attempt to cluster the contents of the ontologies based on the character overlap between the names assigned to roles, classes and instances. Our intent was to use this to estimate how much integration was possible between the different ontologies. The result of our analysis was that while there are a number of distinct integration opportunities, in the great majority of cases they involved a very specific area of the ontology.

Specifically, we discovered a few small ontologies that were very similar and easy to integrate, describing web services and transaction processes. However, we were unable to locate relationships that we had previously discovered through manual querying of the web interface. Our explanation for this is that in the case of the larger biological databases, integration can only be done at very specific points which occur with less relative frequency than in smaller ontologies. Furthermore, we found it disturbing that over 75% of collected ontologies were un-parseable by most of our tools, including Racer [10]. Clearly, much work on robustness remains to be done for both ontological tools and the ontologies themselves.

In the future, we aim to query ontologies with more sophisticated Racer syntax that will reveal ancestor, parent and child concepts and the roles in which their instances participate. This will allow us to perform more detailed ontology compatibility studies and drive future research towards better quality ontology integration methods.

Conclusion

We conclude in saying that ontologies available across the web are not at full maturity. A greater population and larger diversity is needed to in order for a semantically enabled web to be established (The Semantic Web). Our observations of ontology development and release to the community, do however indicate that we are fast approaching a milestone where sufficient numbers of distributed ontologies will be available for ontological integration.

What remains to be determined is the ease of ontology integration across the web and how to evaluate ontologies for the compatibility and integration of their concepts. Domain knowledge, content overlap and common subsets of concepts roles and instances have to be the primary criteria for ontology integration. Additional considerations may include the correctness of axioms, the presence of instance data, the number of properties/roles, class constraints on concepts, the source trustworthiness, ontology freshness, the breadth and depth of taxonomy in the ontology and the suitability of the ontology for inference studies.

Acknowledgments

This work was supported in part by the Fungal Web project, “Ontology, the Semantic Web, Intelligent Systems for Fungal Genomics” (V. Haarslev and G. Butler) funded by Genome Quebec.