# *OntoSearch: Retrieval and Reuse of Ontologies*

*Edward Thomas*, Yi Zhang, Derek Sleeman, Alun Preece, Craig McKenzie, Joe Wright*

*Department of Computing Science, University of Aberdeen, UK*

**ABSTRACT**

This document provides an update on the development of OntoSearch[1][2], an ontology search engine designed to help users find RDF based ontological information on the Semantic Web. It uses the Google API to search several million documents on the Semantic Web and uses these results to populate a local repository of ontological data. This is then searched using queries optimised for the relationships within Ontologies. It supports various visualisation and representational algorithms. These facilities allow the user to make a rapid assessment of the files retrieved.

## 1 OVERVIEW AND MOTIVATION

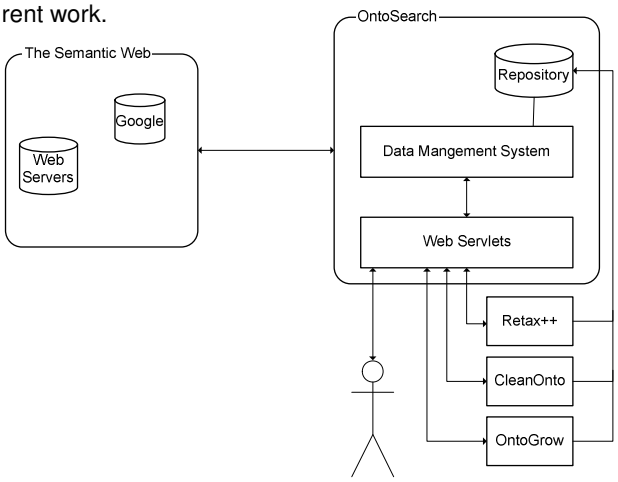
Finding a suitable ontology from the Internet is a hard task because of the difficulty of separating ontological data from the mass of instance data on the Semantic Web[1] and quickly evaluating its suitability.

There is still no good tool to handle this problem. Google offers a powerful web search engine. However, with regard to ontology searching, it has its own problems, such as a lack of visualisation facilities and the poor summary information delivered in search results. Swoogle[2] provides a focussed search of Ontologies on the semantic web, searching for specific keywords appearing as class or property names, but if the keywords only appear in instance data, or as a comment then no match is made.

An opportunity was identified for a tool which provides the breadth of search possible through Google, along with additional functionality to help users interpret these results. OntoSearch has been available in this form for 8 months, and during a recent review session, we obtained the following requirements from some users:

- The ability to specify the type of the file(s) to be returned (OWL, RDF, all)
- The ability to specify the type of entities to be matched by each keyword (concept, attribute, values, comments, all)
- The ability to specify partial or exact matches on entities. So in partial match mode CHEMICAL would match CHEMICALS, CHEMICAL_AGENTS, etc; and of course in exact matching mode only CHEMICAL would be matched.
- The ability to specify a sub-graph to be searched for. For example, concept Animal with concept Pig within 3 links; concepts with particular attributes would be a further variant.
- The ability to visualize each occurrence of a matched entity in a file. So if CHEMICAL was reported as found in a file, then each occurrence of these would be reported systematically, under the user's control. That is the user, through a GUI could specify that the first occurrence of CHEMICAL should be displayed, the second, and so on.

Several of the above features are dependent for their (effective) implementation on a (local) repository of ontologies / ontological fragments. This was the motivation for the current work.




Because we plan to enhance OntoSearch to do extensive analyses of the retrieved ontological files, it seems sensible to create a local repository (and only search the Web again if the local copy is passed its "sell by date"). Some of the additional services which will be available within the context of the Repository will be:

- RETAX++ which is able to check for the well-formed ness of an ontology/taxonomy, spot "concept cycles", and import inconsistent and fragmented ontologies. RETAX++'s main strength is that it helps the user remove the detected inconsistencies by offering the user a range of options [3].

* To whom correspondence should be addressed.

[1] Semantic Web: http://www.w3c.org/2001/sw

[2] Swoogle: http://www.swoogle.org/

- CleanOnto provides another level of checking as to whether an Ontology is conceptually well formed, and essentially addresses the same task as OntoClean [4]. However, we believe that CleanOnto will be easier for domain experts and knowledge engineers to use than OntoClean which uses a fairly esoteric classification system for concepts.
- OntoGrow can help a user grow a new ontology from ontological fragments which are available in the repository.

Additionally, we have plans to make each of the above systems (namely, OntoSearch, ReTax++, CleanOnto & OntoGrow) into Web services; and we expect that users will call on these services as they think necessary. Finally, there is a plan to make OntoSearch into a Protégé Tab [5] which will access the OntoSearch Semantic Web Service described above.

## 2 BACKGROUND TECHNOLOGIES AND RELATED WORK

OntoSearch's repository uses a triple store based around a Berkeley DB Database[3], this has been optimized to allow fast searching for ontologies and ontological relationships which allows us to perform real time searches on large data sets. Each triple is indexed both on the Subject, Predicate, Object values individually but also on combinations of these to allow fast retrieval of data where Subject and Predicate are known (for example searching for a class with a specific name).

This repository is populated during each search with the top 100 results from Google searching for semantic web data only using the keywords submitted in the search page. Any pages already retrieved will be checked that they are less than 7 days old and if not will be refreshed from the original source. Once the repository has been updated, the search is carried out on the local data set using query functions.

## 3 PROGRESS

Currently, OntoSearch is available in two different versions, the "Classic" OntoSearch which uses the Google engine to find keywords in Ontologies, and the more advanced "Alpha" version of the new system which uses Google to populate a local repository of ontological data which can then be searched for specific keywords in specific places in the Ontology. As the local repository grows, these searches become more effective, and the lag time which the system currently suffers, associated with downloading and storing the first 100 results from the Google search in the repository is reduced.

[3] Berkeley DB Database: http://www.sleepycat.com

Work on the alpha version is currently ongoing, this work is mainly aimed at optimising the queries on the repository and making sure that the data is stored in the most efficient manner for retrieval.

## 4 FUTURE WORK

As the repository of ontological data grows, this gives us a larger database to conduct queries of this data. Yi Zhang is currently working on applying query refinement techniques to ontology searching, which will help the user to clearly specify his/her knowledge requirements and further to express them into advanced queries. This work will be integrated with OntoSearch.

The current keyword interface will be supplemented with a query builder to allow users to work with this query language and build effective queries. There will also be a more advanced visualisation system to allow the specific fragments of an ontology which match a query to be highlighted and explored in more detail than is currently possible through the OntoSearch visualisation tool. We will also expand the API to make this fully compliant with the W3C Web Service definition, allowing easier integration of this system with other tools.

In order to make OntoSearch truly independent we plan, at the next development milestone, to replace Google with a semantic web spider, which can update the data in the repository automatically and also search the Internet for new Semantic Web data for a prespecified number of domains. This will remove a current major bottleneck to performance in the system and allow searches to be conducted in real time, removing the delay when content in the repository is updated while a search is taking place.

## ACKNOWLEDGEMENTS

This work is supported under the Advanced Knowledge Technologies (AKT) IRC (EPSRC grant no. GR/N15764/01) comprising Universities of Aberdeen, Edinburgh, Sheffield, Southampton and the Open University. For further information see: http://www.aktors.org

## REFERENCES

1. Zhang Y, Vasconcelos W, and Sleeman D. OntoSearch: An Ontology Search Engine (2004). The Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge.
2. Thomas E, Zhang Y, Sleeman D, Preece A, McKenzie C and Wright J. (2005) OntoSearch: a Service to Support the Reuse

of Ontologies. Demos and Posters of the 2nd European Semantic Web Conference (ESWC 2005).

3. Lam S, Sleeman D and Vasconcelos W. ReTAX+: A Cooperative Taxonomy Revision Tool (2004). The Twenty-fourth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge.
4. Guarino, N and Welty, CA (2004) An overview of OntoClean. In Handbook on Ontologies (eds S Staab & R Studer). Publ: Heidelberg: Springer, pp151-171.
5. Mark A. Musen,., Ray W. Fergerson, Natalya Fridman Noy, and Monica Crubézy. (2002) Protégé-2000: A Plug-in Architecture to Support Knowledge Acquisition, Knowledge Visualization, and the Semantic Web. Stanford Medical Informatics, Stanford University School of Medicine.