# Ontology Search with the OntoSelect Ontology Library

**Paul Buitelaar, Thomas Eigner**

* DFKI GmbH - Language Technology Lab
Saarbrücken, Germany
paulb@dfki.de

**Abstract**

OntoSelect is a dynamic web-based ontology library that harvests, analyzes and organizes ontologies published on the Semantic Web. OntoSelect allows searching as well as browsing of ontologies according to size (number of classes, properties), representation format (DAML, RDFS, OWL), connectedness (score over the number of included and referring ontologies) and human languages used for class- and object property-labels. Ontology search in OntoSelect is based on a combined measure of *coverage*, *structure* and *connectedness*. Further, and in contrast to other ontology search engines, OntoSelect provides ontology search based on a complete web document instead of one or more keywords only.

## 1. Introduction

Most of current work in ontology-based semantic annotation assumes the use of ontologies that are developed specifically for the task at hand. Instead, a more realistic approach would be to access an ontology library and to select one or more appropriate ontologies. Although the large-scale development and publishing of ontologies is still only in a beginning phase, many are already available. To select the most appropriate ontology (or a combination of complementary ontologies) will therefore be an increasingly important subtask in semantic annotation. In recent years web-based ontology libraries and ontology search engines like OntoKhoj (Patel et al., 2003), OntoSelect (Buitelaar, et al., 2004), SWOOGLE (Ding et al., 2004) and Watson (d'Aquin et al., 2007) have been developed to enable this.

## 2. The OntoSelect Ontology Library

OntoSelect[1] is a dynamic web-based ontology library that harvests, analyzes and organizes ontologies published on the Semantic Web. OntoSelect allows searching as well as browsing of ontologies according to size (number of classes, properties), representation format (DAML, RDFS, OWL), connectedness (score over the number of included and referring ontologies) and human languages used for class- and object property-labels.

### 2.1 Collecting and Analyzing Ontologies

OntoSelect uses the Google API to find published ontologies on the web in the following formats: DAML, OWL and RDFS. In the case of OWL, OntoSelect also determines its type (Full, DL, Lite) and indexes this information accordingly. Each class and object property defined by the ontology is indexed with reference to the ontology in which it occurs. Correspondingly, each label is indexed with reference to the corresponding ontology, class or object property, the human language of the label (if available), and a normalized label name, e.g. `TaxiDriver` is normalized to "`taxi driver`". Object properties are handled similarly as classes except that also information on their type (functional, transitive, symmetric) is indexed. Finally, a separate index is build up in which we keep track of the distribution of labels over all collected ontologies. In this way, a ranked list of frequently used labels can be maintained and browsed by the user. The OntoSelect library can be browsed (see Figure 1) by:

*ontology name*
: derived from `owl:Ontology` or URL

*format*
: derived from the ontology URL

*human language*
: derived from `rdfs:label`

*class or property label*
: derived from `rdfs:label`

*included ontologies*
: derived from `owl:imports`

### 2.2 Statistics on Formats and Multilinguality

OntoSelect currently contains over 1400 ontologies. An important aspect to keep track of is the knowledge representation format used for defining these ontologies: DAML, RDFS or OWL. Table 1 gives an overview of the distribution of these formats over the collected ontologies so far. It is interesting to see that the OWL format already shows a clear advance over the other two formats. Tables 2 and 3 give an overview of the distribution of human languages used in the definition of labels for classes and properties – by individual language and by language combination with English. The advance of English over other languages is not surprising as most ontologies still originate from English speaking countries although some start to appear with labels also in other languages, e.g. German and French.

[1] OntoSelect can be accessed at: http://olp.dfki.de/OntoSelect/

**Table 1: Percentage of ontologies by format**

| Format | OWL | DAML | RDFS | Unknown |
|---|---|---|---|---|
| ***Percentage*** | *790* | *262* | *248* | *120* |

**Table 2: Percentage of ontologies with labels in a particular language**

| Language | ***Percentage*** |
|---|---|
| English | *71.0 %* |
| German | *11.5 %* |
| French | *6.0 %* |
| Spanish | *3.7 %* |
| Portuguese | *3.2 %* |
| other | *4.6 %* |

**Table 3: Number of ontologies with labels in English and another language**

| English and other Language Combination | ***Number of Ontologies*** |
|---|---|
| English - German | *18* |
| English - French | *6* |
| English - Portuguese | *6* |
| English - Spanish | *3* |

## 3. Ontology Search

As the Semantic Web continues to grow in terms of developed and published ontologies, it becomes much easier to *find* rather than *construct* an appropriate ontology for a particular application. On the other hand, as more and more ontologies become available to choose from it is correspondingly hard to find the best ontology. The ontology search problem is therefore a very recent topic of research, which only originated with the growing availability of ontologies on the web. A web-based ontology, defined by use of standard Semantic web representation languages such as RDFS and OWL, is in many respects just another web document that can be indexed, stored and retrieved. On the other hand, an ontology is a highly structured document with possibly explicit semantic links to other ontologies. The OntoSelect approach is based on both observations by ranking ontologies by *coverage*, i.e. the overlap between query terms and index terms; by *structure*, i.e. the ratio of class vs. property definitions; and by *connectedness*, i.e. the level of integration between ontologies.

Other approaches have similarly stressed the importance of such measures, e.g. (Alani et al. 2006) describe the "Class Match", "Density", "Semantic Similarity" and "Betweenness" measures. The Class Match and Density measures correspond roughly to our *coverage* and *structure* measure, whereas the Semantic Similarity and Betweenness measure the semantic weight of query terms relative to the different ontologies that are to be ranked. These last two measures are based on the assumption that ontologies are well-structured with equal semantic balance throughout all constitutive parts, which unfortunately is only seldom the case. Another set of measures or rather criteria for ontology ranking and selection has been proposed by (Sabou et al. 2006).

### 3.1 Ontology Search in OntoSelect

Ontology search in OntoSelect is based on a combined measure of *coverage*, *structure* and *connectedness* of ontologies as discussed above. Further, and in contrast to all of the other approaches mentioned above, OntoSelect provides ontology search based on a complete web document instead of one or more keywords only. Obviously this allows for a much more fine-grained ontology search process. For a given document as ontology search query, OntoSelect first extracts all textual data and analyses this with linguistic tools (part-of-speech tagging, morphological analysis) to extract all nouns in the text as these can be expected to represent ontology classes rather than verbs, adjectives or other word classes. To calculate the relevance of each of the available ontologies in OntoSelect, the set of extracted nouns is used to compute three scores (*coverage*, *structure*, *connectedness*) and a combined score as follows:

**Coverage** of an ontology relative to a query document measures how many of the extracted nouns (keywords) in the query document overlap with the set of class labels in the ontology. For this purpose, the nouns are ranked according to statistical relevance ($X^2$), comparing expected frequencies for extracted nouns in a general document collection (Reference Corpus) with those that are observed in the query document.

**Connectedness** measures how much the ontology is connected to other ontologies, i.e. how many ontologies are included and how well these are established relative to other ontologies. As not all included ontologies are valid ontology files, the connectedness measure includes a normalization over the proportion of valid ontology files.

**Structure** measures how detailed the knowledge structure is that the ontology represents. The structure score is based on the observation that more advanced ontologies generally have a large number of properties. Therefore, a relatively large number of properties would indicate a highly structured and hence more advanced ontology. Structure is measured by the number of properties relative to the number of classes in the ontology:

**Combined Score:** As we observed big differences between the ranges of the three scores, we decided to normalize them, i.e. the values of each score are divided by the maximum of all values over all ontologies. Further, because of different levels of importance, the scores have been assigned weights.

### 3.2 An Example of Ontology Search in OntoSelect

The application of the ranking and search algorithm discussed above can be illustrated with an example of ontology search on the topic 'genetics', which may be represented by the Wikipedia page http://en.wikipedia.org/wiki/Gene The results of the keyword extraction and ontology ranking process for this query document are reported by OntoSelect in two tables, one that shows the top 20 keywords extracted from the query document and one with the ranked list of best matching ontologies according to the computed score (see

Figure 2). Combined and individual scores – connectedness, structure, coverage – are shown as well as the matching labels/keywords and their relevance scores. Retrieved and top ranked ontologies include a large number that are indeed of relevance to the 'genetics' topic, e.g. "nciOncology", "bioGoldStandard", "mygrid", "sequence", etc.

Current work on OntoSelect is concerned with the evaluation of the search algorithm and comparing performance with other ontology search engines available. For this purpose we are constructing an evaluation benchmark, consisting of a controlled set of topics with ontologies assigned to these manually. A description of the benchmark and some preliminary results have been reported in (Buitelaar and Eigner, 2007).

## 4. Towards Evaluation of Ontology Search

In order to test the accuracy of our approach we are currently designing an evaluation experiment with a specifically constructed benchmark of 57 ontologies from the OntoSelect library that were manually assigned to 15 different topics represented by one or more Wikipedia pages. In this way we are able to define ontology search as a regular information retrieval task, for which we can give relevance assessments (manual assignment of ontology documents to Wikipedia-based topics) and compute precision and recall for a set of queries (Wikipedia pages).

The benchmark consists of 15Wikipedia topics and 57 out of 1056 ontologies that have been collected through OntoSelect. The 15 Wikipedia topics covered by the evaluation benchmark were selected out of the set of all class/property labels in OntoSelect - 37284 in total - by the following steps:

- Filtering out labels that did not correspond to a Wikipedia page - this left us with 5658 labels (i.e. topic candidates)
- Next, the 5658 labels were used as search terms to filter out labels that returned less than 10 ontologies (out of the 1056 in OntoSelect) - this left us with 3084 labels / topics
- We then manually decided which of these 3084 labels actually expressed a useful topic, e.g. we left out very short labels ('v') and very abstract ones ('thing') – this left us with 50 topics
- Finally, out of these 50 we randomly selected 15 for which we manually checked the ontologies retrieved from OntoSelect - in this step we checked 269 ontologies out of which 57 were judged as appropriate for the corresponding topic

The resulting 15Wikipedia topics with the number of appropriately assigned ontologies are: Atmosphere (2), Biology (11), City (3), Communication (10), Economy (1), Infrastructure (2), Institution (1), Math (3), Military (5), Newspaper (2), Oil (0), Production (1), Publication (6), Railroad (1), Tourism (9). In future work we will report on evaluation results obtained with this benchmark.

## Acknowledgements

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors.

## References

Alani H., Brewster Ch. And Shadbolt N. *Ranking Ontologies with AKTiveRank.* 5th International Semantic Web Conference, Athens GA, USA. 2006.

Buitelaar P., Eigner T. and Declerck Th. *OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection* Demo Session at the 3rd International Semantic Web Conference, Hiroshima, Japan. 2004.

Buitelaar P., Eigner T. *Evaluating Ontology Search* 5th International EON (Evaluation of Ontologies and Ontology-based tools) Workshop at the 6th International Semantic Web Conference, Busan, South-Korea. 2007.

Ding L., Finin T., Joshi A., Pan R., Cost R. S., Peng Y., Reddivari P., Doshi V. C. and Sachs J. *Swoogle: A search and metadata engine for the semantic web.* 13th ACM Conference on Information and Knowledge Management. 2004.

Patel, Ch., Supekar K., Lee Y. and Park, E. K. *OntoKhoj: a semantic web portal for ontology searching, ranking and classification.* 5th ACM international workshop on Web information and data management, ACM Press. 2003. 58-61

Sabou M., Lopez V. and Motta E. *Ontology Selection on the Real Semantic Web: How to Cover the Queens Birthday Dinner?* EKAW 2006 - 15th International Conference on Knowledge Engineering and Knowledge Management. 2006.

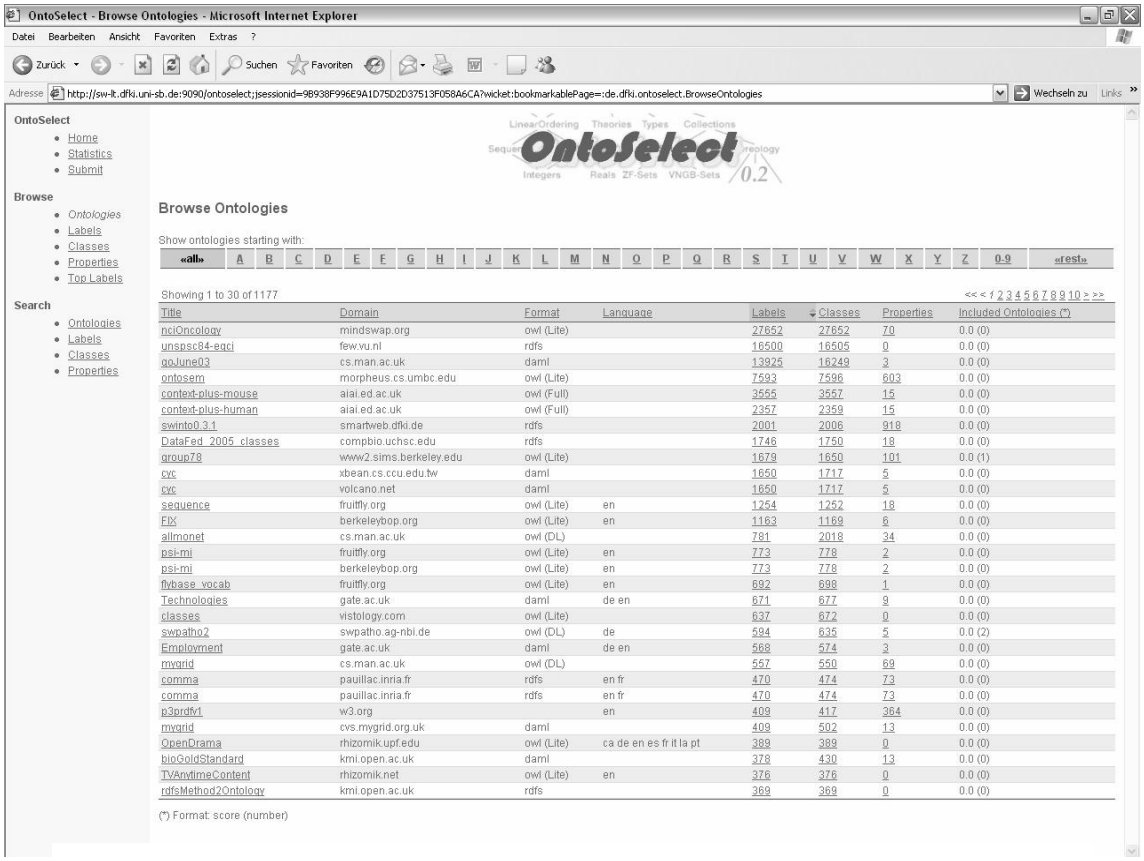


**Figure 1 Browsing ontologies in OntoSelect**

| Score | Title | Matches (*) | Domain | Format | Language | Labels | Classes | Properties | Connectedness | Structure | Coverage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.33 | nciOncology | gene (11643.05), protein (506.9), sequence (359.89), total (336.67), code (241.85), cell (93.68), biology (79.12) | mindswap.org | owl (Lite) | | 27652 | 27652 | 70 | 0.0 | 0.0 | 0.95 |
| 3.0 | bioGoldStandard | gene (11643.05), organism (730.2), process (632.49), structure (372.81), sequence (359.89), diagram (180.47) | kmi.open.ac.uk | daml | | 378 | 430 | 13 | 0.0 | 0.0 | 1.0 |
| 3.0 | mygrid | gene (11643.05), organism (730.2), process (632.49), structure (372.81), sequence (359.89), diagram (180.47) | cvs.mygrid.org.uk | daml | | 409 | 502 | 13 | 0.0 | 0.0 | 1.0 |
| 3.0 | mygrid | gene (11643.05), organism (730.2), process (632.49), structure (372.81), sequence (359.89), diagram (180.47) | cs.man.ac.uk | owl (DL) | | 557 | 550 | 69 | 0.0 | 0.0 | 1.0 |
| 0.87 | sequence | gene (11643.05), protein (506.9) | fruitfly.org | owl (Lite) | en | 1254 | 1252 | 18 | 0.0 | 0.0 | 0.87 |
| 0.87 | psi-mi | gene (11643.05), protein (506.9) | fruitfly.org | owl (Lite) | en | 773 | 778 | 2 | 0.0 | 0.0 | 0.87 |
| 0.87 | psi-mi | gene (11643.05), protein (506.9) | berkeleybop.org | owl (Lite) | en | 773 | 778 | 2 | 0.0 | 0.0 | 0.87 |
| 0.86 | swinto0.3.1 | molecule (1989.74), organism (730.2), protein (506.9), sequence (359.89), expression (331.25), cell (93.68) | smartweb.dfki.de | rdfs | | 2001 | 2006 | 918 | 0.0 | 0.0 | 0.29 |
| 0.36 | umlssn | organism (730.2) | swpatho.ag-nbi.de | owl (DL) | de en | 75 | 87 | 65 | 1.0 | 0.0 | 0.05 |
| 0.24 | goJune03 | transcription (1597.03), protein (506.9), cell (93.68) | cs.man.ac.uk | daml | | 13925 | 16249 | 3 | 0.0 | 0.0 | 0.16 |
| 0.19 | spro_qtv | process (632.49) | www-sop.inria.fr | rdfs | en | 2 | 3 | 6 | 0.0 | 1.0 | 0.05 |
| 0.19 | loggerhead_nesting | organism (730.2), process (632.49), sequence (359.89) | fruitfly.org | owl (Lite) | en | 308 | 314 | 4 | 0.0 | 0.0 | 0.12 |
| 0.16 | gold | transcription (1597.03), process (632.49) | coli.lili.uni-bielefeld.de | owl (Lite) | en | 115 | 117 | 7 | 0.0 | 0.0 | 0.16 |
| 0.11 | w6 | process (632.49) | dannyayers.com | owl (DL) | en | 7 | 8 | 8 | 0.0 | 0.5 | 0.05 |
| 0.1 | dolce2.0-lite-v3 | organism (730.2), process (632.49) | coli.lili.uni-bielefeld.de | owl (DL) | | 81 | 79 | 75 | 0.0 | 0.0 | 0.1 |
| 0.1 | context-plus-human | organism (730.2), process (632.49) | aiai.ed.ac.uk | owl (Full) | | 2357 | 2359 | 15 | 0.0 | 0.0 | 0.1 |
| 0.1 | OBI | organism (730.2), process (632.49) | fugo.sourceforge.net | owl (Full) | en | 153 | 161 | 9 | 0.0 | 0.0 | 0.1 |
| 0.1 | obi | organism (730.2), process (632.49) | berkeleybop.org | owl (Full) | en | 198 | 211 | 15 | 0.0 | 0.0 | 0.1 |
| 0.1 | context-plus-mouse | organism (730.2), process (632.49) | aiai.ed.ac.uk | owl (Full) | | 3555 | 3557 | 15 | 0.0 | 0.0 | 0.1 |
| 0.09 | e3value | diagram (180.47) | e3value.few.vu.nl | rdfs | en | 24 | 29 | 50 | 0.0 | 0.5 | 0.01 |

**Figure 2: Ran ked list of retrieved ontologies for topic ‘genetics’ (Gene)**