CONTEXTUALIZING ONTOLOGIES WITH ONTOLIGHT: A PRAGMATIC APPROACH

Marko Grobelnik, Janez Brank, Blaž Fortuna, Igor Mozetič

Department of Knowledge Technologies, Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
e-mail: {marko.grobelnik, janez.brank, blaz.fortuna, igor.mozetic}@ijs.si

ABSTRACT

We present a pragmatic approach to using large-scale ontologies as contexts. The approach is based on a light-weight ontology model and grounding of the ontology concepts in textual documents. These assumptions allow for efficient implementation of the basic operations (classification, population and mappings between ontologies), and, as a consequence, exploitation of several large-scale ontologies as background, contextual knowledge. We demonstrate one possible scenario how contextual information can be exploited during semi-automatic ontology construction from text corpora.

1 INTRODUCTION

Ontologies represent isolated pieces of knowledge. By networking them, one can explore their interrelations. One form of networked ontologies are contextualized ontologies. In this case, one ontology represents a context of the other and its constituent ingredients (concepts and relations). So, for a given ontology, its ingredients can be interpreted in different contexts by selecting appropriate ontologies which represent appropriate contexts.

In this paper we describe OntoLight which implements basic reasoning functionalities for contextualized ontologies. It is limited to light-weight ontologies which are grounded with appropriate text corpora. The representation and reasoning scales to the largest currently available ontologies, comprising up to one million concepts. In particular, OntoLight currently incorporates the following five ontologies: AgroVoc and ASFA (relevant for the Food and Agricultural Organization of UN), EuroVoc (EU legislation), Cyc (common-sense knowledge) and DMoz (WWW directory).

There are two basic reasoning mechanisms implemented in OntoLight. First, new textual instances without known class can be classified into the selected ontology. Second, soft (probabilistic) mappings between a pair of selected ontologies can be computed, thus providing contextual relationship between the ontologies.

We are using OntoLight as a basic building block for extensions to OntoGen [1], where contextual mappings are used to improve semi-automatic construction of light-weight ontologies from text corpora. The same mechanism of contextual reasoning will be used to extend OntoGen to support simultaneous, collaborative development of an ontology. Our soft mappings between grounded ontologies also complement methods for ontology alignment, where mappings are computed on the basis of common, background ontologies (as provided by Swoogle, for example) [2], [3].

The paper presents the software package OntoLight consisting of several executable modules and data library of ontologies. The main functionality we cover is contextualization of ontologies through generation of soft mappings between ontologies, thus enabling to view concepts of one ontology through the perspective of another one. The second goal was achieving scalability needed for large case studies – i.e. being able to deal with large ontologies such as AgroVoc and ASFA. To achieve this we constrained the representation to a light-weight ontology model which covers targeted functionality needed in case studies. Finally, we took care of the software engineering aspects of the result – namely, the software package is built on top of an existing Text-Garden software library [4]. It is written in C++ with proper API and accessible through several development platforms (Java, Python, Matlab, Mathematica, Prolog).

In the next Section we first present the ontology model used in OntoLight. Next, in Section 3 we present the library of ontologies already incorporated in OntoLight – each ontology is presented through its main features. In Section 4, the software package is presented by describing each module separately and through possible integration of the modules which could be used in a pipeline. Finally, in Section 5, we show an integration of OntoLight with OntoGen, where light-weight ontologies are used as background, contextual knowledge which helps the users during the process of semi-automatic ontology construction from text corpora.

2 THE ONTOLOGY MODEL

The ontology model used in OntoLight is a relatively simple model which covers most of the well known light-weight ontologies. The model we use is a subset of richer ontology formalisms (such as OWL) in the sense that richer ontologies could be imported but not all their expressiveness can be used. Informally, the light-weight ontology model is defined by:

- A list of languages used for lexical terms.

- A list of class-types used for representing different types of nodes in the ontology structure.
- A list of classes where each class can have several lexical representations in one or several languages. One class represents one node in the graph.
- A list of relation-types used to label relations (links) between classes in the ontology graph.
- A list of relations connecting classes in the ontology graph.
- Each ontology can have one or several grounding models. Each grounding model is a function which proposes zero, one or more classes for a given instance. This corresponds to a classification /categorization model in machine learning terminology.

The above model has a one-to-one mapping into C++ classes in the OntoLight module of the Text-Garden library [4].

3 LIBRARY OF ONTOLOGIES

To perform experiments on real data, we had to import several ontologies into the OntoLight framework. Since most of the larger real life ontologies are still in non-standard formats we needed to develop specialized filters for pre-processing the available data into the common ".OntoLight" format used by the rest of the OntoLight package. In the first version of the software we decided to prepare filters for importing five medium to large scale ontologies. They are all used on a daily basis in real life applications. They model different types of knowledge – from relatively specific ones (AgroVoc, ASFA), a general one with legal bias (EuroVoc) to generic ones for Web contents (DMoz) and common sense (Cyc).

3.1 AgroVoc

AgroVoc is a multilingual structured thesaurus of all subject fields in Agriculture, Forestry, Fisheries, Food security and related domains (e.g. Sustainable Development, Nutrition, etc). It consists of words or expressions (terms) in different languages and is organized in the thesaurus relationships (e.g. "broader", "narrower", and "related") used to identify or search resources. Its main role is to standardize the indexing process in order to make search simpler and more efficient, and to provide users with the most relevant resources.

The AgroVoc thesaurus was developed by the Food and Agriculture Organization of the United Nations (FAO) and the Commission of the European Communities, in the early 1980s. It is updated by FAO roughly every three months and users can see the specific changes on the AgroVoc website [5]. AgroVoc is available in the five official languages of FAO, which are English, French, Spanish, Chinese and Arabic. Additionally, it is also available in Czech, German, Japanese, Portuguese, Slovak and Thai. Other translations, such as Hindi, Hungarian, Italian and Korean are currently underway or being revised.

AgroVoc is downloadable in several formats – we used the MS Access package which includes several tables with all the data about the ontology. Specifically, AgroVoc includes 12 languages, 65 relation-types, and 47101 classes. AgroVoc classes were grounded with text abstracts from ASFA document corpus (see below) which are close to AgroVoc terms.

3.2 ASFA

ASFA (Aquatic Sciences and Fisheries Abstracts) is a thesaurus used for the Aquatic Sciences and Fisheries Information System (ASFIS), an international co-operative information system for the collection and dissemination of information covering the science, technology and management of marine, brackish water, and freshwater environments. It contains approximately 1 million bibliographic references to the world's aquatic science literature accessioned since 1971 (for some journals and/or subject areas the coverage precedes 1971). All references are machine readable.

ASFA is produced as a cooperative effort by the international network of ASFA partners [6] which consists of: United Nations Co-sponsoring Partners, National and International Partners, and the Publishing Partner. The objective is to disseminate bibliographic information to the relevant research community. A good description of several aspects of ASFA is available at [7].

In our case we extracted the ASFA thesaurus and abstracts by crawling the web search interface. The extracted data were all in the English language. The thesaurus structure included two types of classes (descriptor and non-descriptor), 5 link types, and 9882 classes. ASFA classes were grounded with text abstracts available within the records of the crawled data (over 360.000 abstracts).

3.3 EuroVoc

EuroVoc is a multilingual thesaurus covering the fields in which the European Communities are active – it provides a means of indexing the documents in the documentation syste?ms of the European institutions and of their users. The European Parliament, the Office for Official Publications of the European Communities, the national and regional parliaments in Europe, some national government departments and European organizations are currently using this controlled vocabulary. The recent version EuroVoc 4.2 exists in 21 official languages of the European Union (Bulgarian, Spanish, Czech, Danish, German, Estonian, Greek, English, French, Italian, Latvian, Lithuanian, Hungarian, Dutch, Polish, Portuguese, Romanian, Slovak, Slovene, Finnish and Swedish), and one other language (Croatian). In addition to these versions, it has been translated by the Parliaments of several other countries: Albania, Russia and Ukraine.

The data of the thesaurus are available from [8] where we extracted the thesaurus structure by crawling the html pages

(since the officially proposed way of getting the data was non-functioning) while the multilingual part (without the structure) was downloadable from the web site as an MS Excel file. The extracted data is available in 21 languages, it has two types of nodes (descriptors and non-descriptors), 5 relation types, and 13416 nodes (out of which 6645 are descriptors). We grounded the EuroVoc classes with the documents from Acquis Communitarian, the corpus of European legislation indexed with EuroVoc descriptors.

3.4 Cyc

The Cyc [9] knowledge base (KB) is a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. The original form of representation is a formal language CycL. The KB consists of terms which constitute the vocabulary of CycL and assertions which relate those terms. These assertions include both simple ground facts and rules with variables.

Cyc KB is available for researchers from the Cycorp company homepage [10] in two different forms – OpenCyc (vocabulary only) and ResearchCyc (full version). In our case, we are using the data retrieved directly from the company under the ResearchCyc license. Since Cyc KB is very rich (it includes ~50.000 first order logic rules) we decided to deal only with the static part of the KB. It is written only in English, it has two types of classes (concepts and lexical nodes), it has 3295 relations, and 464.988 concepts.

Since Cyc has only structure (concepts and facts) we grounded each Cyc's concept by querying Google with lexical representation for that class.

3.5 DMoz/Open Directory Project

The Open Directory Project (ODP), also known as DMoz, is the largest multilingual open content directory of World Wide Web links that is constructed and maintained by a community of volunteer editors. The browsing and search service is accessible from [11].

The directory data (structure and content) are available from [12] in the RDF format. The version we are using here uses only the English part of the directory, it has 3 types of relations, and 642.995 concepts.

The taxonomic part was grounded with the content which is available within the downloadable data. The main data source for grounding were short textual descriptions of the manually categorized web sites within each DMoz category.

4 SOFTWARE MODULES

In the following subsections we present each of the OntoLight modules (or module groups) dealing with ontology data – from raw data to classification models and mappings. The software is available from [13].

4.1 Ontology data transformation utilities

The function of the ontology data transformation utilities is to process specific formats of each of the selected ontologies for the ontology library. The result of all the utilities is saving the ontology data in the unifying binary format with the file-extension ".OntoLight" and its textual counterpart with the file extension ".OntoLight.Txt". As described in section 3, the ontology library consists of five ontologies – therefore we prepared five command line utilities for processing the data:

- AgroVoc2OntoLight.Exe
- Asfa2OntoLight.Exe
- Cyc2OntoLight.Exe
- DMoz2OntoLight.Exe
- EuroVoc2OntoLight.Exe

Each of the utilities takes on the input file name or file path to the data and produces binary file (".OntoLight") and textual file (".OntoLight.Txt"). An example run of the transformation of the EuroVoc is the following:

```
[d:\textgarden\eurovoc2ontolight]
EuroVoc2OntoLight.exe
EuroVoc To Ontology-Light [Feb 12 2007]
=======================================
Input-EuroVoc-FilePath (-i:)=f:/data/EuroVoc/
Output-OntoLight-FileName (-
o:)=f:/Data/OntoLight/EuroVoc.OntoLight
Output-Text-FileName (-
ot:)=f:/Data/OntoLight/EuroVoc.OntoLight.Txt
=========================
Loading 'f:/data/EuroVoc/listMultiLg_All.txt'
... 6645/6646
Done. (6645)
Loading 'f:/data/EuroVoc/eurovoc.txt' ...
Done. (48044)
Saving OntoLight to
'f:/Data/OntoLight/EuroVoc.OntoLight' ...
Done.
Saving Text to
'f:/Data/OntoLight/EuroVoc.OntoLight.Txt' ...
Done.
```

4.2 Ontology grounding module

The ontology grounding module OntoLight2OntoCfier.exe creates from an ontology stored in the ".OntoLight" format an additional file with the extension ".OntoCfier" (and its textual representation ".OntoCfier.Txt"). This file includes a classification model which is used by OntoClassify module (next subsection) for classification of new instances in the ontology classes. The current version uses a centroid-based classifier which calculates a centroid vector for each class in the ontology. It takes into account data used for grounding and the hierarchical part of the ontology structure. The actual classification is performed with the kNN (k-nearest-neighbour) algorithm [14].

Here is an example run of the OntoLight2OntoCfier.exe module for ontology grounding. On the input the utility takes ".OntoLight" data and a pre-processed Bag-Of-Words file with the text documents and the descriptors from the ontology. On the output the system creates ".OntoCfier" file with a classifier and its textual representation (".OntoCfier.Txt"). With additional parameters we specify the language we are using for grounding (in the case when data exists in several languages), to see whether the document's category equals descriptors in the ontology and the threshold for writing weighted words in the textual output.

```
[d:\textgarden\ontolight2ontocfier]OntoLight2
OntoCfier.exe
Ontology-Light To Ontology-Classifier [Feb 12
2007]
================================================
======
Input-OntoLight-FileName (-
iol:)=f:/Data/OntoLight/EuroVoc.OntoLight
Input-BagOfWords-FileName (-
ibow:)=f:/Data/OntoLight/Acquis.Bow
Output-OntoClassifier-FileName (-
oom:)=f:/Data/OntoLight/EuroVoc.OntoCfier
Output-OntoClassifier-Text-FileName (-
oom:)=f:/Data/OntoLight/EuroVoc.OntoCf
Language-Name (-lang:)=EN
DocumentCategory-Is-TermId (-catisid:)=Yes
Cut-Word-Weight-Sum-Percent (-cwwprc:)=0.33
====================================
Loading Onto-Light from
'f:/Data/OntoLight/EuroVoc.OntoLight' ...
Done.
Loading Bag-Of-Words from
'f:/Data/OntoLight/Acquis.Bow' ... Done.
Generating Ontology-Classifier...
  Creating BowDocWgtBs ... Done.
  Collecting documents per ontology-term ...
    Docs:7972/7972 Pos:26915 Neg:149
  Done.
  Creating sub-terms & up-terms vectors ...
Done.
  Creating centroids ...
    Active-Terms:1399
    Active-Terms:441
    Active-Terms:85
    Active-Terms:7
    Active-Terms:0
    Active-Terms:0
  Done.
Done.
Saving Onto-Classifier to
'f:/Data/OntoLight/EuroVoc.OntoCfier' ...
Done.
Saving Text to
'f:/Data/OntoLight/EuroVoc.OntoCfier.Txt' ...
Done.
```

4.3 Ontology population module

The ontology population module OntoClassify.Exe takes as input a grounded ontology in the ".OntoCfier" format and instance data (in various textual formats) and produces XML and textual file with the possible categories for the given instance.

In the following example we take a grounded version of the EuroVoc and the query "Slovenia and Croatia are having a fishing industry". The result is in the files OntoCfy.Xml and OntoCfy.Txt.

```
[d:\textgarden\ontoclassify]OntoClassify.exe
Ontology-Classify [Feb 12 2007]
================================
Input-OntoClassifier-FileName (-
ioc:)=f:/Data/OntoLight/EuroVoc.OntoCfier
Input-Query-String (-qs:)=Slovenia and
Croatia are having a fishing industry.
Input-Query-Html-File (-qh:)=
Input-Query-CompactDocument-FileName (-
qcpd:)=
Input-Query-Url (-qu:)=
Input-Query-URL-Vector-FileName (-quf:)=
Output-Classification-Xml-File (-
ox:)=OntoCfy.Xml
Output-Classification-Txt-File (-
ot:)=OntoCfy.Txt
=================
Loading Onto-Classifier from
'f:/Data/OntoLight/EuroVoc.OntoCfier' ...
Done.
```

The resulting textual file lists classes from the EuroVoc grounded ontology to which the query should belong with the highest confidence. Each line of the file OntoCfy.Txt includes the following three fields: rank, confidence, and class name:

```
 1. 0.201 Croatia
 2. 0.171 fisheries policy
 3. 0.162 Slovenia
 4. 0.161 fishing area
 5. 0.159 national independence
 6. 0.159 fishing regulations
 7. 0.156 fishery management
 8. 0.147 fisheries structure
 9. 0.147 fishing fleet
10. 0.144 Community fisheries
```

4.4 Ontology mapping module

The last module in the pipeline of utilities is the utility OntoJoint.exe which takes as an input two grounded ontologies in the ".OntoCfier" format and creates soft mappings between the classes of both ontologies. This is done in the following way: first, by aligning vocabularies of grounded ontologies (this typically means aligning words from respective bag-of-words representations), and second, by classifying centroid vectors from the first ontology into the classes of the second one.

In the following example we take as an input the EuroVoc and ASFA ontologies and store mapping results into XML and textual files, OntoJoint.XML and OntoJoint.Txt, respectively:

```
[d:\textgarden\ontojoint]OntoJoint.exe
```

```
Join-Ontologies [Mar 12 2007]
==============================
Input-OntoClassifier-FileName-1 (-
ioc1:)=f:/Data/OntoLight/EuroVoc.OntoCfier
Input-OntoClassifier-FileName-2 (-
ioc2:)=f:/Data/OntoLight/Asfa.OntoCfier
Output-OntologyJoin-Xml-File (-
ox:)=OntoJoint.Xml
Output-OntologyJoin-Txt-File (-
ot:)=OntoJoint.Txt
===============
Loading Onto-Classifier-1 from
'f:/Data/OntoLight/EuroVoc.OntoCfier' ...
Done.
Loading Onto-Classifier-2 from
'f:/Data/OntoLight/Asfa.OntoCfier' ... Done.
```

The following is an example mapping from the resulting OntoJoint.Txt file where we see a mapping from the ASFA "fishing licence" class to 10 related classes from the EuroVoc ontology.

```
'fishing licence' →
    1. 'Legal aspects' (0.003)
    2. 'Ships' (0.003)
    3. 'Disputes' (0.002)
    4. 'Ecology' (0.002)
    5. 'Military operations' (0.001)
    6. 'Rare species' (0.001)
    7. 'Public health' (0.001)
    8. 'Fish culture' (0.001)
    9. 'Commercial fishing' (0.001)
   10. 'Resource development' (0.001)
```

5 CONTEXUALIZED ONTOLOGY GENERATION WITH OntoGen

OntoGen [1] is a software tool for semi-automatic, data-driven ontology construction. It incorporates methods for discovering concepts from a collection of documents. Documents are represented by the well known bag-of-words representation, where each document is encoded as a vector of term frequencies. The similarity of a pair of documents is calculated by the number and weights of the words that these documents share. The weights of the words are usually calculated by the so called TFIDF weighting, but there are other alternatives.

OntoGen implements two methods for concept discovery: Latent Semantic Indexing (LSI) [15] and k-means clustering [16]. LSI is a method for linear dimensionality reduction by learning an optimal sub-basis which approximates documents' bag-of-words vectors. The sub-basis vectors are proposed as concepts. The k-means method discovers concepts by clustering the documents' bag-of-words vectors into k clusters where each cluster is a proposed concept.

We have extended OntoGen with OntoLight, specifically with five general-purpose light-weight ontologies: AgroVoc, ASFA, EuroVoc, DMoz and Cyc. These ontologies provide contexts to the user during the user-guided, data-driven generation of an ontology from a corpus of documents. OntoGen structures the documents into concepts and subconcepts, but, until now, has used only extracted keywords to suggest concept names. With contextual ontologies available, OntoGen is now able to provide much better suggestions for concept names based on the similarity between structured documents and grounded concepts from the selected contexts. As a consequence, the user can view each concept suggested by OntoGen through different "sematic lenses": each view corresponds to a different context as implemented by a different light-weight ontology. Figure 1 gives an example.

6 CONCLUSION

In the paper we describe OntoLight, a set of software modules for:

- transforming raw ontology data for several ontologies from their specific formats into a unifying light-weight ontology format,
- grounding the ontology and storing it into grounded ontology format,
- populating grounded ontologies with new instance data, and
- creating mappings between grounded ontologies.

As a part of OntoLight we already prepared the ontology library consisting of five different ontologies: AgroVoc, ASFA, Cyc, DMoz, and EuroVoc. Additional ontologies (e.g., WordNet) will be incorporated in the future.

We will be using OntoLight as a basic building block for extensions to OntoGen, where contextual mappings are used to improve semi-automatic construction of light-weight ontologies from text corpora. The same mechanism of contextual reasoning will be used to extend OntoGen to support simultaneous, collaborative development of an ontology. Our soft mappings between grounded ontologies also complement methods for ontology alignment, where mappings are computed on the basis of common, background ontologies. We plan to integrate our approach to mappings with the mechanisms for ontology alignments.

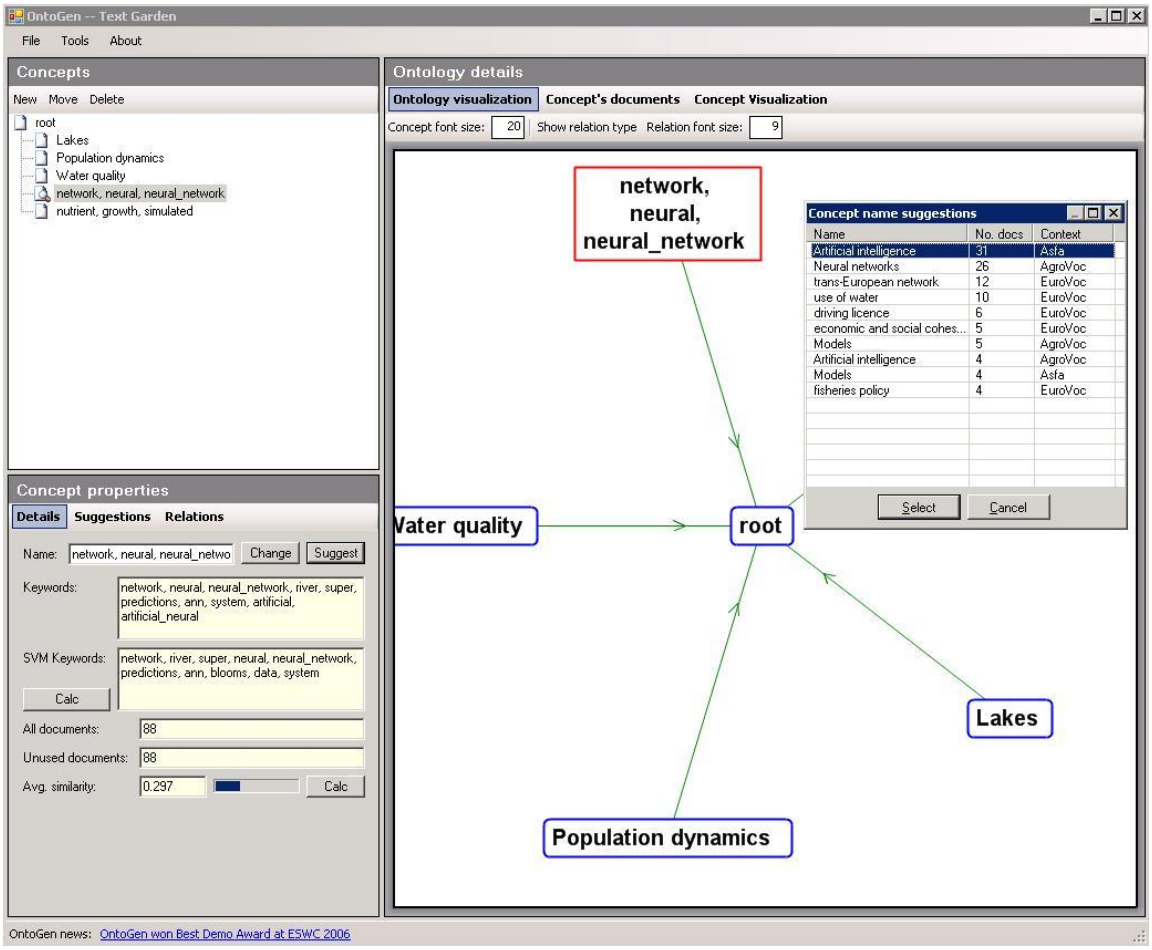

Figure 1: *A screenshot of OntoGen when used to structure the abstracts of recent issues of the Ecological Modelling journal. Contexts are provided by three ontologies: AgroVoc, ASFA, and EuroVoc. Some concept names were already derived from contextual suggestions (Water quality, Population dynamics, Lakes) and the user inspects current suggestions for the top node (network, neural, neural_network). The system provides two sensible suggestions: Artificial Intelligence (from ASFA) and Neural networks (from AgroVoc), while the third suggestion: trans-European network (from EuroVoc) probably makes less sense.*

Acknowledgement

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under NeOn (IST-2004-27595-IP) and PASCAL (IST-2002-506778).

References

1. B. Fortuna, M. Grobelnik, D. Mladenic: Semi-automatic Construction of Topic Ontology. Semantics, Web and Mining, Joint International Workshop, EWMF 2005 and KDO 2005, Porto, Portugal, October 3-7, 2005.
2. M. Sabou, M. d'Aquin, and E. Motta: Using the Semantic Web as Background Knowledge for Ontology Mapping, In Proceedings of the International Workshop on Ontology Matching (OM-2006), collocated with ISWC-06.
3. M. Sabou, M. d'Aquin, W. R. van Hage and E. Motta: Improving Ontology Matching by Dynamically Exploring Online Knowledge. Submitted for review, 2007.
4. TextGarden: http://kt.ijs.si/Dunja/textgarden/
5. AgroVoc: http://www.fao.org/aims/ag_intro.htm
6. ASFA: http://www.fao.org/fi/asfa/partners.asp
7. ASFA: http://www.fao.org/fi/website/FIRetrieveAction.do?dom=org &xml=asfa_prog.xml&xp_nav=2
8. EuroVoc: http://europa.eu/eurovoc/
9. Douglas B. Lenat. Cyc: A Large-Scale Investment in Knowledge Infrastructure. Comm. ACM 38, no. 11, November 1995.
10. Cyc: http://www.cyc.com/
11. DMoz: http://dmoz.org/
12. DMoz: http://rdf.dmoz.org/
13. OntoLight: http://analytics.ijs.si/Projects/NEON/OntoLight.Zip
14. Shakhnarovish, Darrell, Indyk (Eds.): Nearest-Neighbor Methods in Learning and Vision, The MIT Press, 2005.
15. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman: Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, vol. 41, no. 6, 391-407, 1990.
16. Jain, Murty, Flynn: Data Clustering: A Review, ACM Comp. Survey, 1999.