NeOn: Lifecycle Support for Networked Ontologies

Integrated Project (IST-2005-027595)

Priority: IST-2004-2.4.7 – “Semantic-based knowledge and content systems”

D3.2.1 Reasoning with contexts – prototype interpreter

Deliverable Co-ordinator: Marko Grobelnik

Deliverable Co-ordinating Institution: Jozef Stefan Institute (JSI)

Other Authors: Janez Brank (JSI); Igor Mozetic (JSI)

This deliverable provides software prototype OntoLight for contextualization of light-weight ontologies. Within the software package there are several software modules in the form of command line utilities and the library of five large ontologies: AgroVoc, ASFA, EuroVoc, Cyc and DMoz/ODP. Functionalities of the package include (1) processing of different input data formats into a common light-weight ontology format, (2) grounding of the ontologies, (3) population of the ontologies (classification of new instances into ontology classes), (4) creation of contextual, soft mappings between any pair of ontologies.

Document Identifier:	NEON/2007/D3.2.1/v1.0	Date due:	February 28, 2007
Class Deliverable:	NEON EU-IST-2005-027595	Submission date:	March 30, 2007
Project start date:	March 1, 2006	Version:	v1.1
Project duration:	4 years	State:	Final
		Distribution:	Public

NeOn Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities, grant number IST-2005-027595. The following partners are involved in the project:

Open University (OU) – Coordinator Knowledge Media Institute – KMi Berrill Building, Walton Hall Milton Keynes, MK7 6AA United Kingdom Contact person: Martin Dzbor, Enrico Motta E-mail address: {m.dzbor, e.motta} @open.ac.uk	**Universität Karlsruhe – TH (UKARL)** Institut für Angewandte Informatik und Formale Beschreibungsverfahren – AIFB Englerstrasse 28 D-76128 Karlsruhe, Germany Contact person: Peter Haase E-mail address: pha@aifb.uni-karlsruhe.de
Universidad Politécnica de Madrid (UPM) Campus de Montegancedo 28660 Boadilla del Monte Spain Contact person: Asunción Gómez Pérez E-mail address: asun@fi.upm.es	**Software AG (SAG)** Uhlandstrasse 12 64297 Darmstadt Germany Contact person: Walter Waterfeld E-mail address: walter.waterfeld@softwareag.com
Intelligent Software Components S.A. (ISOCO) Calle de Pedro de Valdivia 10 28006 Madrid Spain Contact person: Richard Benjamins E-mail address: rbenjamins@isoco.com	**Institut ‘Jožef Stefan’ (JSI)** Jamova 39 SI-1000 Ljubljana Slovenia Contact person: Marko Grobelnik E-mail address: marko.grobelnik@ijs.si
Institut National de Recherche en Informatique et en Automatique (INRIA) ZIRST – 655 avenue de l'Europe Montbonnot Saint Martin 38334 Saint-Ismier France Contact person: Jérôme Euzenat E-mail address: jerome.euzenat@inrialpes.fr	**University of Sheffield (USFD)** Dept. of Computer Science Regent Court 211 Portobello street S14DP Sheffield United Kingdom Contact person: Hamish Cunningham E-mail address: hamish@dcs.shef.ac.uk
Universität Koblenz-Landau (UKO-LD) Universitätsstrasse 1 56070 Koblenz Germany Contact person: Steffen Staab E-mail address: staab@uni-koblenz.de	**Consiglio Nazionale delle Ricerche (CNR)** Institute of cognitive sciences and technologies Via S. Martino della Battaglia, 44 - 00185 Roma-Lazio, Italy Contact person: Aldo Gangemi E-mail address: aldo.gangemi@istc.cnr.it
Ontoprise GmbH. (ONTO) Amalienbadstr. 36 (Raumfabrik 29) 76227 Karlsruhe Germany Contact person: Jürgen Angele E-mail address: angele@ontoprise.de	**Asociación Española de Comercio Electrónico (AECE)** C/Alcalde Barnils, Avenida Diagonal 437 08036 Barcelona Spain Contact person: Jose Luis Zimmerman E-mail address: jlzimmerman@fecemd.org
Food and Agriculture Organization of the United Nations (FAO) Viale delle Terme di Caracalla 1 00100 Rome, Italy Contact person: Marta Iglesias E-mail address: marta.iglesias@fao.org	**Atos Origin S.A. (ATOS)** Calle de Albarracín, 25 28037 Madrid Spain Contact person: Tomás Pariente Lobo E-mail address: tomas.parientelobo@atosorigin.com

Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed writing parts of this document:

JSI

FAO

Change Log

Version	Date	Amended by	Changes
0.1	2007-03-11	Marko Grobelnik	First release
0.2	2007-03-12	Igor Mozetic	Executive summary, minor modifications
1.0	2007-03-15	Igor Mozetic	Comments by Jerome Euzenat
1.1	2007-03-26	Igor Mozetic	Final QA comments

Executive Summary

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.

This deliverable, OntoLight, 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 FAO case study), EuroVoc (EU legislation), Cyc (common-sense knowledge) and DMoz (WWW directory).

There are two basic reasoning mechanisms implemented in OntoLight. First, new instances can be classified into selected ontology, thus providing appropriate context for the instances. 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, 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 a common, background ontology.

Table of Contents

1. Introduction

This deliverable 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 the NEON 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 the NEON 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. 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 the software package. Next, in Section 3 we present the library of ontologies prepared within the deliverable – each ontology is presented through its main features. Finally, 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.

2. The Ontology Model

The ontology model used in the software package OntoLight uses a relatively simple ontology model which covers most of the well known light-weight ontologies. The model we use is a subset of richer ontology formalisms in the sense that richer ontologies could be imported but not all their expressiveness can be used.

In an informal way, our ontology model can be described in the following way.

An ontology is defined with:

- List of languages used for lexical terms.
- List of class-types used for representing different types of nodes in the ontology structure.
- List of classes where each class can have several lexical representations in one or several languages. One class represents one node in the graph.
- List of relation-types used to label relations (links) between classes in the ontology graph.
- 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.h/.cpp module of Text-Garden library.

3. Library of Ontologies

To perform experiments on the real data, we had to import several ontologies into the OntoLight framework. Since most of the larger real life ontologies are still in the 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 5 medium to large scale ontologies. They are all used on a daily basis in the 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 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 searching 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 http://www.fao.org/aims/ag_intro.htm .

AGROVOC is available in the five official languages of FAO, which are English, French, Spanish, Chinese and Arabic. It is also available in Czech, German, Japanese, Portuguese, Slovak and Thai. Other languages such as Hindi, Hungarian, Italian and Korean are currently either being translated or revised.

AgroVoc is downloadable in several formats – we used 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) being 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 (http://www.fao.org/fi/asfa/partners.asp) 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 the following web site: http://www.fao.org/fi/website/FIRetrieveAction.do?dom=org&xml=asfa_prog.xml&xp_nav=2 .

In our case we extracted the ASFA thesaurus and abstracts by crawling of 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.

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 systems 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 organisations 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 the web site http://europa.eu/eurovoc/ 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 EuroVoc classes with the documents from Acquis Communitarian, the corpus with European legislation being indexed with EuroVoc descriptors.

3.4 Cyc

The Cyc 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 http://www.cyc.com/ 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 ResearchCyc licence. Since Cyc KB is very rich (it includes ~50000 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 464988 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 http://dmoz.org/ .

The directory data (structure and content) are available from http://rdf.dmoz.org/ 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 642995 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 will present each of the modules (or module groups) dealing with ontology data – from raw data to classification models and mappings. The software is available from the URL http://analytics.ijs.si/Projects/NEON/OntoLight.Zip .

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"). 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

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 kNN (k-nearest-neighbour) algorithm.

Here is the example run of the OntoLight2OntoCfier.exe module for ontology grounding. On the input the utility takes input ".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 case the 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]OntoLight2OntoCfier.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.OntoCfier.Txt
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

Ontology population module OntoClassify.Exe takes on the input grounded ontology in ".OntoCfier" format and an 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 comes 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, 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
11. 0.144 fishing grounds
12. 0.143 common fisheries policy
13. 0.140 fishing vessel
14. 0.138 catch quota
15. 0.138 catch of fish
16. 0.137 fishing controls
17. 0.133 authorised catch
18. 0.133 fishing permit
19. 0.128 fishing agreement
20. 0.128 conservation of resources
21. 0.116 conservation of fish stocks
```

```
22. 0.111 financial compensation of an agreement
23. 0.111 fishing rights
24. 0.100 fish
25. 0.098 territories of the former Yugoslavia
```

4.4 Ontology mediation module

The last module in the pipeline of utilities is the utility OntoJoint.exe which takes on the 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 one ontology into the classes of the second one.

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

```
[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.
```

An example mapping from the resulting OntoJoint.Txt file is the following where we see 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. Conclusion

In the deliverable we created a set of software modules for:

- transforming raw ontology data for several ontologies from their specific formats into unifying 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 the software package we prepared the ontology library consisting of five different ontologies: AgroVoc, ASFA, Cyc, DMoz, and EuroVoc.

The software is available from the URL http://analytics.ijs.si/Projects/NEON/OntoLight.Zip .

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 a common, background ontology. We plan to integrate our approach to mappings with the mechanisms for ontology alignments.