NeOn-project.org

NeOn: Lifecycle Support for Networked Ontologies

Integrated Project (IST-2005-027595)

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

D3.2.5 Cyc plug-in for NeOn Toolkit

Deliverable Co-ordinator: Dunja Mladenić

Deliverable Co-ordinating Institution: J. Stefan Institute (JSI)

Other Authors: Luka Bradeško (JSI), Lorand Dali (JSI), Blaž Fortuna (JSI), Marko Grobelnik (JSI), Inna Novalija (JSI), Boštjan Pajntar (JSI)

This deliverable provides prototype of a Cyc plug-in for NeOn Toolkit that enables answering question given in natural language based on a collection of documents. The user provides a collection of documents from the domain of interest that will be used for answering questions. Cyc ontology together with any other domain specific ontology is used to provide semantic information in the form of synonyms and generalisations of the terms that occur in the documents. AnswerArt technology is used for answering questions provided in natural language.

Finally, a NeOn Toolkit plug-in was implemented to provide answers to the questions using Cyc and provide contextualized visual representation of the results.

The prototype is demonstrated on ASFA abstracts using ASFA thesaurus and WordNet in addition to Cyc, all in English.

Document Identifier:	NEON/2009/D3.2.5/v1.0	Date due:	October 31st, 2009
Class Deliverable:	NEON EU-IST-2005-027595	Submission date:	October 31st, 2009
Project start date:	March 1, 2006	Version:	V1.0
Project duration:	4 years	State:	Final
		Distribution:	Public

NeOn Consortium

This document is a part of the NeOn research project funded by the IST Programme of the Commission of the European Communities by the 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 11 D-76128 Karlsruhe, Germany Contact person: Andreas Harth E-mail address: aha@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: Jesús Contreras E-mail address: jcontreras@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	**Food and Agriculture Organization of the United Nations (FAO)** Viale delle Terme di Caracalla 1 00100 Rome Italy Contact person: Caterina Caracciolo E-mail address: Caterina.Caracciolo@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	**Laboratorios KIN, S.A. (KIN)** C/Ciudad de Granada, 123 08018 Barcelona Spain Contact person: Antonio López E-mail address: alopez@kin.es

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 to the writing of this document or its parts:

JSI

Change Log

Version	Date	Amended by	Changes
0.1	9-10-2009	Dunja Mladenić	Overall structure of the report
0.2	20-10-2009	Marko Grobelnik	Data preparation
0.3	02-11-2009	Inna Novalija	Data analysis
0.4	4-11-2009	Lorand Dali	Data enhancement
0.5	6-11-2009	Luka Bradeško, Lorand Dali, Inna Novalija	Adjustment of Cyc
0.6	9-11-2009	Blaž Fortuna	Connecting Cyc with AnswerArt
0.7	10-11-2009	Boštjan Pajntar	NeOn Toolkit plug-in
0.8	12-11-2009	Dunja Mladenić	Architecture diagram
0.9	13-11-2009	Marko Grobelnik	Overall revision
0.10	16-11-2009	Lorand Dali	Usage description
0.11	18-11-2009	Boštjan Pajntar	Example usage and Showcase
0.12	19-11-2009	Dunja Mladenić	Executive summary, Introduction, Discussion, Overall revision
0.13	27-11-2009	Dunja Mladenić	Revision based on the review (1-4)
0.14	3-12-2009	Boštjan Pajntar	Revision based on the review (5-10)

Executive Summary

This report describes a software deliverable developed as an integration of Cyc ontology [Lenat, 1995] and AnswerArt technology [Dali et al., 2009] implemented as a plug-in for NeOn Toolkit. The prototype provides contextualized answers to questions provided in natural language. The contextualization is achieved by using Cyc ontology and possibly some additional domain specific ontology. The answers are provided based on a domain specific document collection.

The approach consists of several phases as follows. In the data preparation phase, we extract the relevant part of Cyc ontology and extend it with any other relevant ontology either general or domain specific. In this phase, the document collection is pre-processed using AnswerArt technology to obtain subject-predicate-object triplets from the sentences in the domain specific document collection. In the data enhancement phase, the triplets are enhanced using semantic knowledge obtained from the extended part of Cyc ontology. In the data indexing phase, the enhanced triplets are index for efficient search for answers. In the final phase, the question is transformed based on predefined patterns of questions to enable efficient search over the indexed triplets and the list of answers is returned.

This software deliverable is implemented as a NeOn Toolkit plug-in giving contextualized answers to the questions provided in natural language.

The prototype is demonstrated in English on domains specific document collection being ASFA abstracts, using ASFA thesaurus and WordNet as domain specific ontologies in addition to Cyc ontology. The approach is applicable to other natural languages, assuming the necessary natural language processing is handled and the match between the document collection language, domain ontology language and Cyc ontology is ensured.

Table of Contents

1. Introduction 6

2. Motivation 7

3. Approach Description 8

4. Underlying technology 9

4.1 AnswerArt description 9

4.2 Cyc description 10

4.3 NeOn Toolkit plug-in 10

5. Data description – ASFA 12

6. Data enhancement 14

6.1 Data enhancement with Cyc 14

6.2 Data enhancement using Wordnet 14

6.3 Data enhancement using ASFA 15

7. Example usage of the system 16

7.1 Showcase 16

8. Discussion and future work 21

List of Figures

Figure 1. Architecture of the approach. 8

Figure 2. Illustration of the AnswerArt system functionality. 9

Figure 3: Starting the CYC AnswerArt Plug-in 16

Figure 4: Initial view of the plug-in 17

Figure 5: Returned answers 17

Figure 6: Detailed description of the plug-in 18

Figure 7: Document View 19

Figure 8: Summary of the document 19

Figure 9: Extended summary of the document 20

1. Introduction

This deliverable presents the system for contextualized access to the data by providing answers to questions formulated in natural language. The technology behind is based on AnswerArt [Dali et al., 2009] technology for question answering and Cyc [Lenta, 1995] for providing semantic context to the document collection from a particular domain of interest.

In the first step, the document collection is pre-processed using natural language processing to identify subject-predicate-object triplets. In the second step Cyc ontology is extended by any additional domain specific ontologies. We show application of the proposed approach on ASFA abstracts using extension of Cyc by WordNet as a general ontology and ASFA ontology as a domain specific ontology.

The proposed approach is implemented in a prototype that uses service oriented architecture connecting all the components in a plug-in for NeOn Toolkit.

The NeOn Toolkit plug-in can be found at: http://kameleon.ijs.si/cyc/. Both the binary and source version of the plug-in are available. The web application can be found at: http://fox.ijs.si/proxy/answerart/neon/.

The remaining part of this deliverable is structured as follows. Section

2. Motivation provides a brief description of the motivation behind this work. Section 3. Approach Description describes the approach giving an overview of the architecture. In Section 4. Underlying technology we briefly describe the underlying AnswerArt technology. Section

The plug-in requires eclipse version 3.2 and Java runtime version of 1.5 or higher which coincides with the toolkit requirements.

5. Data description – ASFA gives description of the ASFA data that was used as a basis for answering questions, while the details on the extension of Cyc ontology using WordNet and ASFA ontology are in Section 6. Data enhancement. Usage of the system is described in Section 7. Example usage of the system. Section 8. Discussion and future work concludes with a short discussion and some ideas for future development.

2. Motivation

The proposed approach is based on combining AnswerArt [Dali et al., 2009] technology for question answering based on triplets and Cyc [Lenta, 1995] ontology for providing semantic context to the document collection from a particular domain of interest. When performing search, we are frequently looking for answers to some specific questions, the proposed approach enables obtaining answers to questions provided in natural language. Moreover, the approach provides answers in semantic context of the domain of interest via including Cyc ontology extended by any other domain specific ontology as needed.

The Cyc ontology is a formalized representation of a large amount of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. In addition to impressive amount of common sense knowledge, Cyc contains domain specific knowledge organized in so called "microtheories". Moreover, Cyc has already been successfully used for semantic consolidation of triplets extracted from natural language using similar technology as in AnswerArt [Baxter et al., 2009] for ontology generation and document summarization.

AnswerArt integrates two important functionalities: providing answers to questions and browsing through the document that supports the answer. The questions follow a predetermined template, whereas the answers are yielded based on the previously extracted information, in the form of subject – predicate – object triplets. Furthermore, the system retrieves the sentences that support these answers, as well as the documents containing the sentences. It integrates three possibilities of further exploring the relevant documents, which provide a document overview: by analyzing the list of facts (subject – verb – object triplets) extracted from the document, by visualizing the semantic representation of the document and by browsing the document summary. Related approaches query structured data stored in ontologies, while AnswerArt derives the answers only from unstructured text, which means that the things the user can ask about are not limited or domain specific. TextRunner [Banko and Etizioni, 2008] is similar to AnswerArt in the way that it also involves applying structured queries on unstructured text, while the main difference is that AnswerArt also provides a natural language interface to the search. The Calais[1] system creates semantic metadata for user submitted documents in the form of named entities, facts and events. On the other or AnswerArt named entities and facts represent the starting point and they are further refined by applying co-reference resolution for named entities, anaphora resolution and semantic normalization based on WordNet for facts. This process enables the construction of a semantic description of the document in the form of a semantic directed graph where the nodes are the subject and object triplet elements, and the link between them is determined by the predicate. Powerset[2] enables search over Wikipedia and Freebase, where the search results contain aggregated information from several articles, as well as a list of facts related to people, places and things. The main difference is that AnswerArt describe the answer by a visual representation of the document in the form of a semantic graph and by the document summary, which is automatically extracted based on the document semantic graph.

[1] Calais web page: http://www.opencalais.com/

[2] Powerset web page: http://www.powerset.com/

3. Approach Description

We propose an approach that enables answering question from a desired domain using a collection of relevant documents. The approach consists of four phases: data preparation, data enhancement, data indexing and question handling. Architecture of the proposed approach is given in Figure 1.

In the data preparation phase, we extract the relevant part of Cyc ontology and extend it with any other relevant ontology either general or domain specific. In our application scenario we have selected ASFA abstracts for a document collection and extended Cyc by using WordNet and ASFA ontology. The document collection is pre-processed using AnswerArt technology to obtain subject-predicate-object triplets out of the sentences in the document collection.

In the data enhancement phase, the triplets are enhanced using semantic knowledge obtained from the extended Cyc ontology. In particular, each part of the triplet is extended by a set of synonyms and direct generalizations obtained from the ontology.

In the data indexing phase, the extended triplets are index for efficient search for answers. The search is performed by transforming each question into a set of semi-triplets – triplets with missing one or two arguments. Search is performed as matching of the semi-triplets with the triplets from the index in order to find possible values of the missing arguments (answers to the question).

In the final phase, the question is transformed based on predefined patterns of questions to enable efficient search over the indexed triplets and the list of answers is returned.

In the question handling phase, the question is transformed based on predefined patterns of questions to enable efficient search over the indexed triplets and the list of answers is returned.

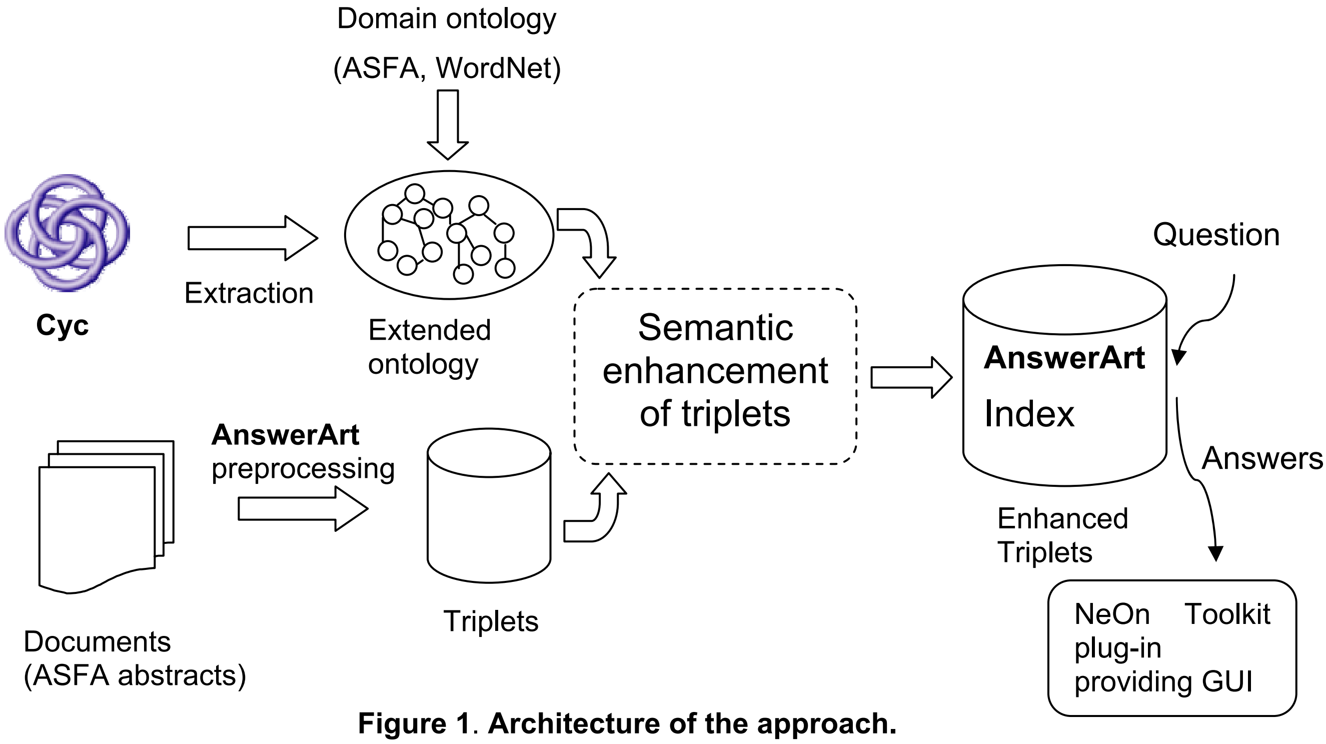


Figure 1. Architecture of the approach.

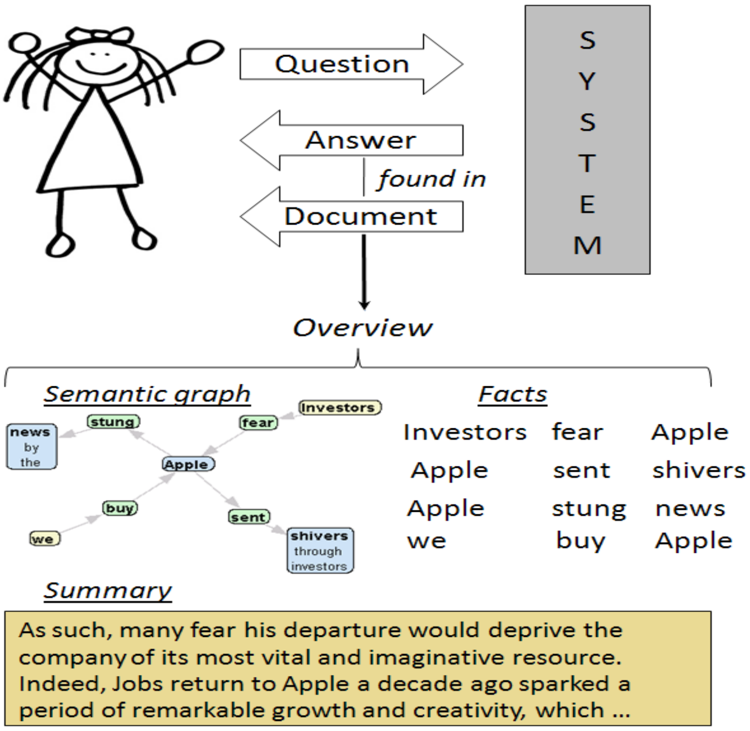
4. Underlying technology

The proposed approach builds on several existing technologies. Components of the AnswerArt system for question answering are adjusted for including semantic enhancement of the data using Cyc ontology.

4.1 AnswerArt description

AnswerArt [Dali et al., 2009] combines question answering, summarization and document visualization functionalities. The user obtains answers based on the facts previously extracted from text in the form of subject – predicate – object triplets. Moreover, the sentences that support the answer, as well as the documents containing these sentences, are also retrieved. The relevant documents can be further explored with the aid of a document overview functionality that consists of a document summary, a semantic representation of the initial document and a list of facts extracted from the document.

The system searches for possible answers to the question and, when found, each answer is linked to the sentences that support it and the document that contains these sentences. The system provides a document overview by retrieving the document semantic graph, the list of subject – predicate – object facts and the automatically generated document summary of variable length that is set interactively by the user.



Investors	fear	Apple
Apple	sent	shivers
Apple	stung	news
we	buy	Apple

As such, many fear his departure would deprive the company of its most vital and imaginative resource. Indeed, Jobs return to Apple a decade ago sparked a period of remarkable growth and creativity, which ...

Figure 2. Illustration of the AnswerArt system functionality.

Extraction of subject – predicate – object facts is a pre-processing step to document collection. Triplets are extracted from each sentence in turn. This means that to extract triplets from a document, the text in that document has to be split into sentences. Moreover, each sentence is tokenized and the tokens are tagged with their part of speech. After this, chunking is performed.

Chunking means that several related consecutive tokens are grouped together, based on their tags, resulting in phrases (noun phrases and verb phrases), also called chunks. Having chunked a sentence, simple rules can be applied to extract triplets from it. An example of such a rule would be: a noun phrase followed by a verb phrase followed by another noun phrase is a triplet.

4.2 Cyc description

Cyc Knowledge Server is a very large, multi-contextual knowledge base and inference engine, developed for more than twenty years with a goal to break the "software brittleness bottleneck" once and for all by constructing a foundation of basic "common sense" knowledge--a semantic substratum of terms, rules, and relations. This enables a variety of knowledge-intensive products and services. Cyc is intended to provide a "deep" layer of understanding that can be used by other programs to make them more flexible. The Cyc technology includes the following components:

- The Cyc Knowledge Base
- The Cyc Inference Engine
- The CycL Representation Language
- The Natural Language Processing Subsystem
- Cyc Semantic Integration Bus
- Cyc Developer Toolsets

For the purposes of the Cyc Neon plug-in we only used the Knowledge base, Inference Engine and The Natural Language Processing Subsystem, which we accessed through Cyc Developer Toolsets (CYC API).

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 medium of representation is the formal language CycL, described below. The KB consists of terms--which constitute the vocabulary of CycL--and assertions which relate those terms. These assertions include both simple ground assertions and rules.

The Cyc KB is divided into "microtheories" (currently thousands of). Assertions in each of the microtheories share a common set of assumptions; some microtheories are focused on a particular domain of knowledge, a particular level of detail, a particular interval in time, etc. The microtheory mechanism allows Cyc to independently maintain assertions which are prima facie contradictory, and enhances the performance of the Cyc system by focusing the inferencing process.

At the present time, the Cyc KB contains nearly two hundred thousand terms and several dozen hand-entered assertions about/involving each term. New assertions are continually added to the KB by human knowledge enterers and lately also with the help of the machines using various machine learning algorithms. Additionally, term-denoting functions allow for the automatic creation of millions of non-atomic terms, such as (LiquidFn Nitrogen); and Cyc adds a vast number of assertions to the KB by itself as a product of the inferencing process.

4.3 NeOn Toolkit plug-in

In order to provide the proposed technology inside the NeOn toolkit as a plug-in, service oriented architecture was used. The proposed approach consists of several modules. First, there is a module for the triplet extraction, which can work over any document repository. In our implementation we demonstrate it over the ASFA abstracts. Next, there is a module which consolidates Cyc and domain ontologies, which all provide semantic enhancements of the triple set. Last there is AnswerArt module for translating natural language questions into triplet queries and visualization and summarization of the results. All of these modules run on different servers, so the most rational approach was to use service oriented architecture.

We have integrated the online AnswerArt application inside the NeOn Toolkit as a plug-in. The prerequisite is of course a working internet connection, which enables communication between all of the modules. The end result is a NeOn Toolkit plug-in, which enables the user to access the application without being aware of the complex architecture that runs in the background.

The developed plug-in can be obtained from http://kameleon.ijs.si/cyc/ in both the binary and source code. The web application which is integrated inside the plug-in can be found at: http://fox.ijs.si/proxy/answerart/neon/.

To install the plug-in it is necessary to copy the binary version inside the plug-ins folder of the NeOn toolkit installation. No additional plug-ins are required. Toolkit needs to be restarted before the first use.

The plug-in requires eclipse version 3.2 and Java runtime version of 1.5 or higher which coincides with the toolkit requirements.

5. Data description – ASFA

Aquatic Sciences and Fisheries Abstracts (ASFA) [Fagetti et al., 2009] is a database covering the literature on the science, technology, management, and conservation of marine, brackish water, and freshwater resources and environments, including their socio-economic and legal aspects.

More than 5,000 serial publications, books, reports, conference proceedings, translations and limited distribution literature are selected for abstracting and indexing in ASFA. Publications in more than 40 languages, with English as primary language, are represented in ASFA database that actually aggregates five databases:

ASFA-1: Biological Sciences and Living Resources,

ASFA-2: Ocean Technology, Policy and Non-Living Resources,

ASFA-3: Aquatic Pollution and Environment Quality,

ASFA Aquaculture Abstracts,

ASFA Marine Biotechnology Abstracts.

ASFA database contains 367696 records; a typical record is described with the following fields:

TI:	Title
AU:	Author
AF:	Author Affiliation
SO:	Source
IS:	ISSN
AB:	Abstract
LA:	Language
SL:	Summary Language
PY:	Publication Year
PT:	Publication Type
DE:	Descriptors
ER:	Environmental Regime
TR:	ASFA Input Center Number
CL:	Classification
SF:	Subfile
AN:	Accession Number

We have pre-processed the ASFA documents using AnswerArt technology. There were 3100832

triplets extracted from the data containing 347 403 unique terms. Examples of the extracted triplets are the following (with [cyc], [wn-…], [asfa-…] indicating the source of semantic information Cyc, Wordnet and ASFA respectively):

Salmon	spawn	rivers
Sockeye	salmon [cyc]	migrated week
Salmon	is	export product
disease	affecting	tissues
disease	cured	antibiotics
relation	disease	development
symptoms	associated	disease
weight	caught	survey

ship	deform	ice
ship	has	ballast tanks
fisheries	conservation	areas
blood	developing	mechanism
cDNA[wn-gen]	encoding	alpha sub
heart	using	leucine
Olyutorsk-Navarin population	has	centers
species	described	middle
wave spectra	obtained	HF radar
phytoplankton	represented	species
carbon	fixation	increased %
sunlight[wn-gen|cyc]	supported	growth

Using synonymic, hypernymic relations for WordNet and synonymic, hierarchical relations for Cyc and ASFA thesaurus, we have extracted concepts, correspondent to the unique terms from ASFA documents. The number of types of inferences for correspondent Cyc, WordNet and ASFA extracted terms is given below:

[cyc]: 228 266	Cyc relations
[wn-gen]: 4 2942	WordNet hypernymic relations
[wn-syn]: 37 777	WordNet synonymic relations
[asfa-gen]: 4 214	ASFA hierarchical relations
[wn-syn|cyc]: 3 839	WordNet synonymic/Cyc relations
[asfa-syn]: 3 245	ASFA synonymic relations
[wn-gen|cyc]: 1 976	WordNet hypernymic/Cyc relations
[asfa-gen|cyc]: 104	ASFA hierarchical/Cyc relations
[wn-syn|wn-gen]: 102	WordNet hypernymic/WordNet synonymic relations
[wn-syn|asfa-syn]: 89	WordNet synonymic/ASFA synonymic relations
[asfa-syn|cyc]: 58	ASFA synonymic/Cyc relations
[wn-gen|asfa-gen]: 55	WordNet hypernymic/ASFA hierarchical relations
[wn-syn|asfa-syn|cyc]: 33	WordNet synonymic/ASFA synonymic relations
[wn-syn|wn-gen|cyc]: 13	WordNet synonymic/WordNet hypernymic/Cyc relations
[wn-gen|asfa-syn]: 9	WordNet hypernymic/ASFA synonymic relations
[wn-gen|asfa-gen|cyc]: 4	WordNet hypernymic/ASFA hierarchical/Cyc relations
[wn-syn|asfa-gen]: 2	WordNet synonymic/ASFA hierarchical relations
[wn-gen|asfa-syn|cyc]: 2	WordNet hypernymic/ASFA synonymic relations
[wn-syn|asfa-gen|cyc]: 1	WordNet synonymic/ASFA hierarchical/Cyc relations

6. Data enhancement

6.1 Data enhancement with Cyc

The input for this part of the task was the subject-predicate-object triplets extracted from ASFA documents. The triplet source was then connected into the Cyc KB using OpenCYC API.

For each from the vast amount of concepts (347,404) in English the Cyc KB was queried to get the corresponding Cyc concept (see the examples below). Furthermore, for each concept that Cyc have its English representation we queried it few times, to get the related concepts, especially its generalizations. For each of the concept we usually get one or more of its generalized meanings and then the Cyc was queried again to get the English presentations of all related concepts. This English was then stored for later use in the question answering.

Out of the 347403 concepts occuring in the extracted triplets, 10310 are covered by cyc. There are 228266 inferences made by cyc, which means that one concept has roughly about 20 related concepts extracted from cyc.

Example:

marine is related to: **marine personnel, military person, military persons, military people, military personnel, serviceman, servicemen, most military, more military, military personnel**

6.2 Data enhancement using Wordnet

WordNet is a lexical database of the English language containing about 150 000 words. These words are organized into synsets. Each synset is made of a set of words which have the same meaning, and also contains an explanation, examples, and the part of speech which the words in that synset have. A synset can have sematic relations to other synsets. Two of these relations are the hypernymy relation and the hyponymy relations. These two are relevant for understanding the described system.. Synset X is sayd to be a hypernym of synsset Y if every Y is also an X (e.g. animal is a hypernym of dog), and synset X is a hyponym of Y if every X is also a Y (e.g. greyhound is a hyponym of dog). In other words hypernyms are more general forms and hyponyms are more specific. The whole WordNet database is organized into a hierarchy of hypernyms and hyponyms.

The system uses WordNet to find related terms for each concept extracted from the ASFA abstracts. By related term we mean: synonyms (the other words in the synset) andone level of hypernyms (more general words which are one level up in the WordNet hierarchyThe goal of the related terms is to improve the recall of the search for a given concept, which will be found not only if the user searches for the word as it appeared in the text, but also if he searches for a related term.

There were extracted 347404 terms from ASFA, for 27775 of which synonyms could be found in wordnet, and for 20358 hypernyms could be found.

Example:

catfish synonyms: **mudcat**

hypernyms: **freshwater fish**

6.3 Data enhancement using ASFA

ASFA thesaurus [CSA ASFA Database Guide] contains over 9800 concepts, applies to database indexing and provides a set of terms used by indexers to describe the contents of publications. These thesaurus terms are listed in the Descriptors field of each record in ASFA database.

All thesaurus terms are either Descriptors or Non-descriptors. Descriptors or allowable (permitted) terms represent concepts used in ASFA database indexing and searching.

Non-descriptors or forbidden (or unauthorized) terms include true synonyms, quasi-synonyms, word forms, different (American) spelling or very specific terms grouped for indexing (or retrieval) purposes into a conceptually broader term.

In ASFA thesaurus Descriptors and Non-descriptors are connected through USE and Used For (UF) fields. Non-descriptors are followed by the USE reference which leads to the relevant descriptor and UF references in the descriptor section list all non-descriptors for the particular descriptor.

Each concept might contain the information about hierarchical and affinitive relations with other thesaurus concepts. Generic hierarchical Broad Term (BT) and Narrow Term (NT) relations represent superclass and subclass taxonomic relationships. Non-hierarchical Related Term (RT) relations may indicate antinomy, suggest possible concurrent use of two concepts or indicate other than hierarchical relationships.

Scope Note (SN) represents the definition of the scope of the term, provides additional information and usage history.

4948 ASFA thesaurus concepts corresponded to 347403 ASFA abstract terms have been extracted and elaborated with synonymic and hierarchical relations.

For each mapped ASFA thesaurus - ASFA abstract concept USE, UF and 2-level BT related concepts have been found:

```
<DESCRIPTOR>Oil pollution</DESCRIPTOR>              Oil pollution →
   <BT>Pollution</BT>                                           Pollution


<DESCRIPTOR>Disease resistance</DESCRIPTOR>         Disease resistance →
  <UF>Disease susceptibility</UF>                               Disease susceptibility
  <UF>Pathogen resistance</UF>                                  Pathogen resistance
  <UF>Resistance to disease</UF>                                Resistance to disease
  <BT>Biological resistance</BT>                                Biological resistance
                                                                Biological properties
<DESCRIPTOR>Biological resistance</DESCRIPTOR>
  <BT>Biological properties</BT>
```

7. Example usage of the system

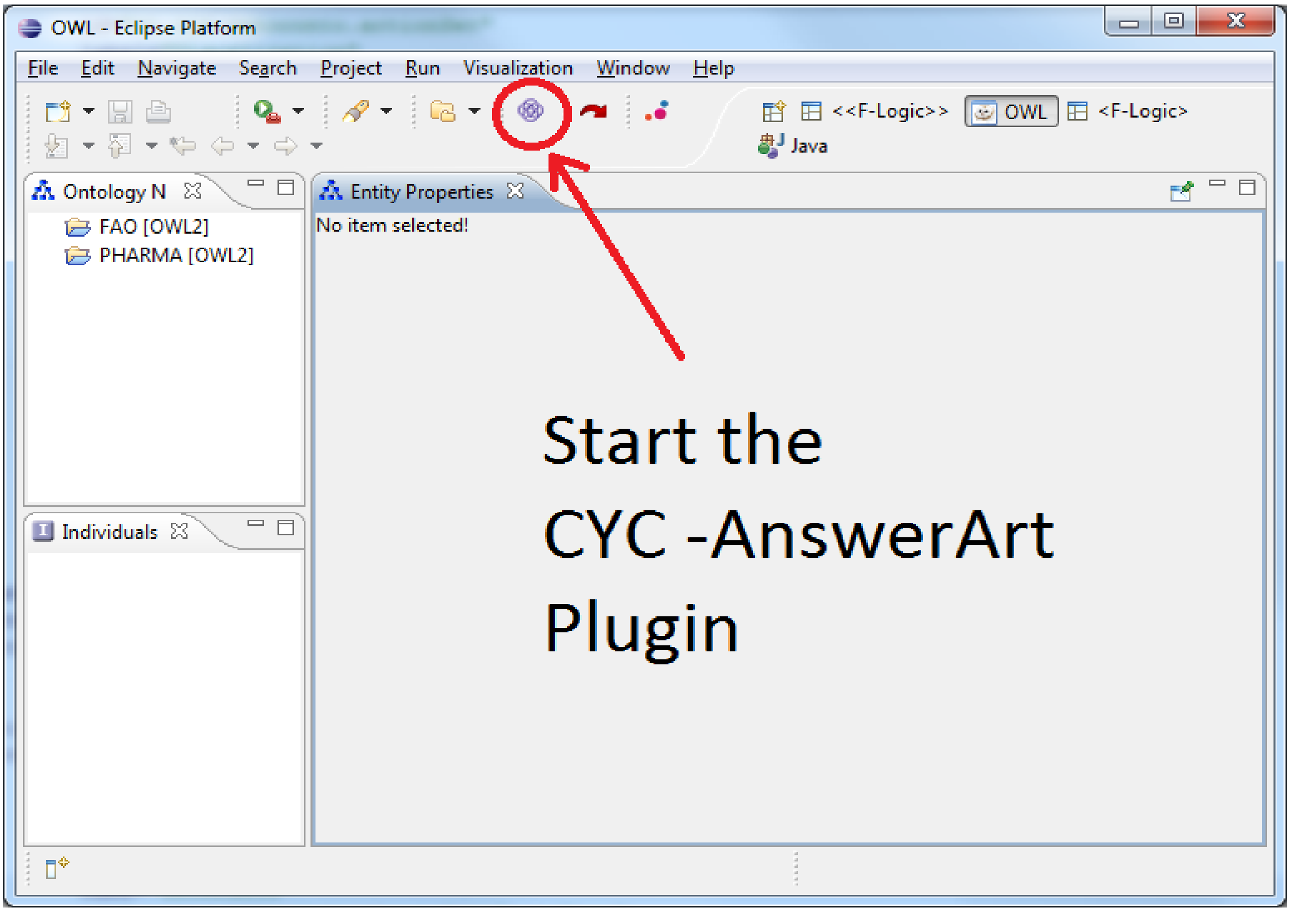
The main goal of the developed NeOn Toolkit plug-in is to provide question answering functionality in contextualized way. In general we derive answers from a repository of textual documents in the form of triples. These triples get semantically enhanced using Cyc Knowledge Server (Section 3.2, 5.1) and other domain ontologies. This also provides contextualization of knowledge extraction, since different ontologies enhance different triples. The AnswerArt module provides functionality of question answering, transforming natural language questions into triples queries, visualizing and summarizing the results (Section 3.1). Over this basic setting we have implemented a concrete application.

To demonstrate the technology, we have decided to have an example of FAO case study and implement a NeOn toolkit plug-in that uses ASFA abstracts as the document repository. Besides using Cyc, we have additionally enhanced the results by using the ASFA thesaurus, which provides domain specific context and WordNet ontology for additional general context.

7.1 Showcase

Here we provide a showcase of an actual usage of the plug-in. We demonstrate all the possible user interaction within the plug-in, and explain results.

To start the plug-in, the easiest way is to press the Cyc toolbar button (Figure 3). It is also possible to start from the menu Visualization -> Cyc AnswerArt, or from Window -> Show View -> Other -> Visualization -> Cyc AnswerArt.



No item selected!

Figure 3: Starting the CYC AnswerArt Plug-in

After the plug-in is loaded, there is an empty input field where a natural language question can be posed. Next to it is a button "Ask in ASFA abstracts" to submit (Figure 4). In our example we pose a question: "What could pollution have affected" and results are shown in Figure 5., with specific elements of the GUI marked in Figure 6.

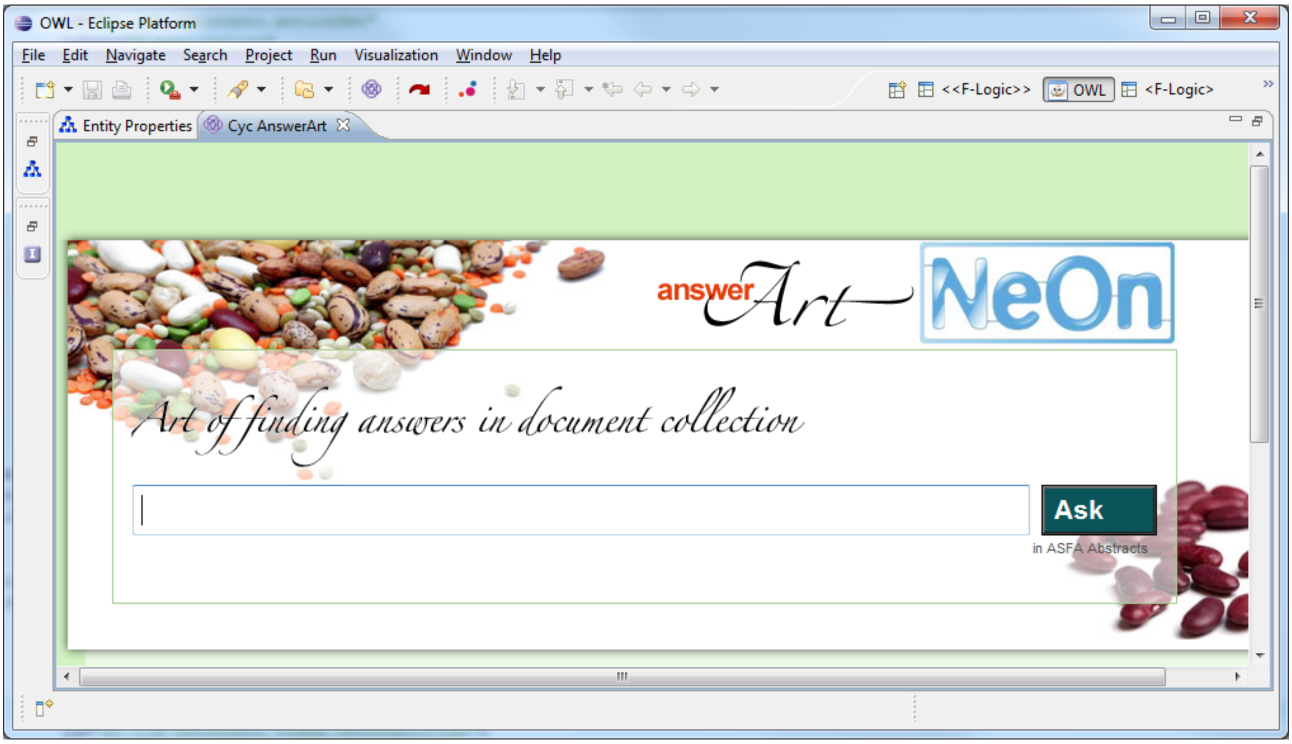
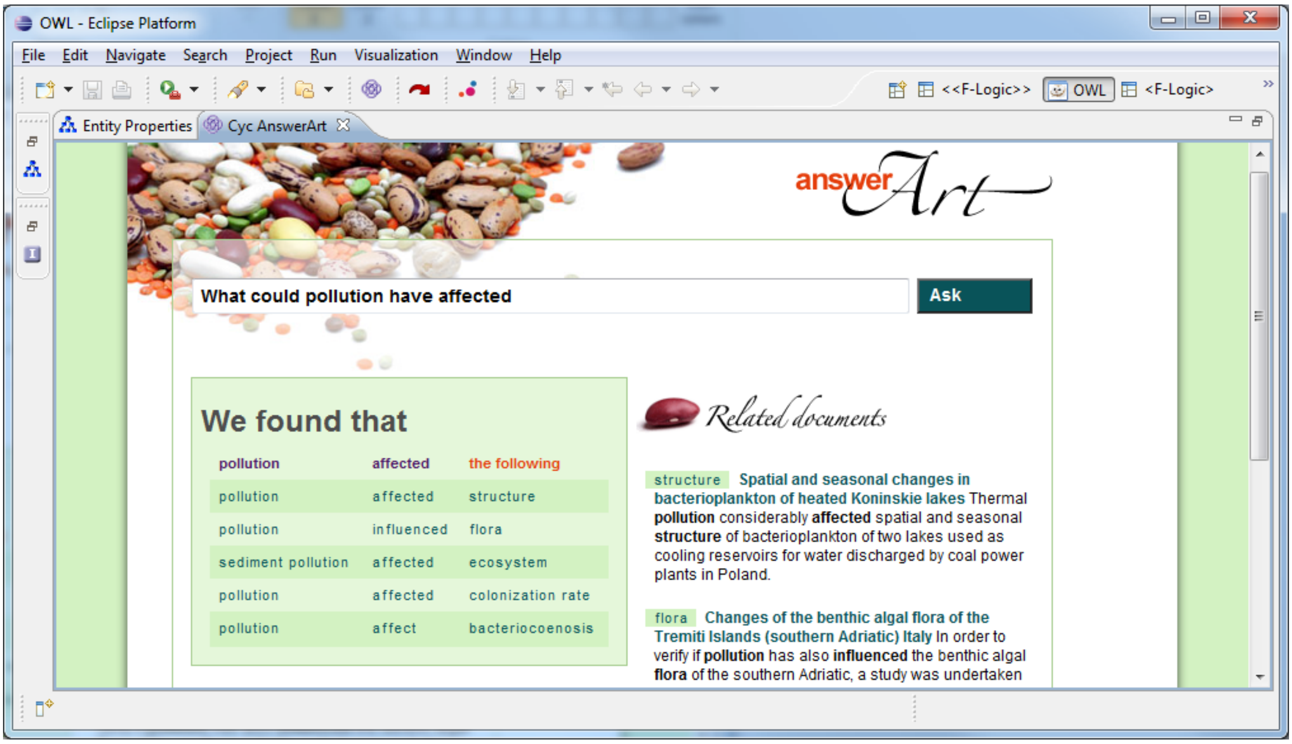

OWL - Eclipse Platform
File Edit Navigate Search Project Run Visualization Window Help
<<F-Logic>> OWL <F-Logic>
Entity Properties
Cyc AnswerArt
answerArt
NeOn
Art of finding answers in document collection
Ask
in ASFA Abstracts

Figure 4: Initial view of the plug-in


OWL - Eclipse Platform
File Edit Navigate Search Project Run Visualization Window Help
<<F-Logic>> OWL <F-Logic>
Entity Properties
Cyc AnswerArt
answerArt
Ask

What could pollution have affected

We found that

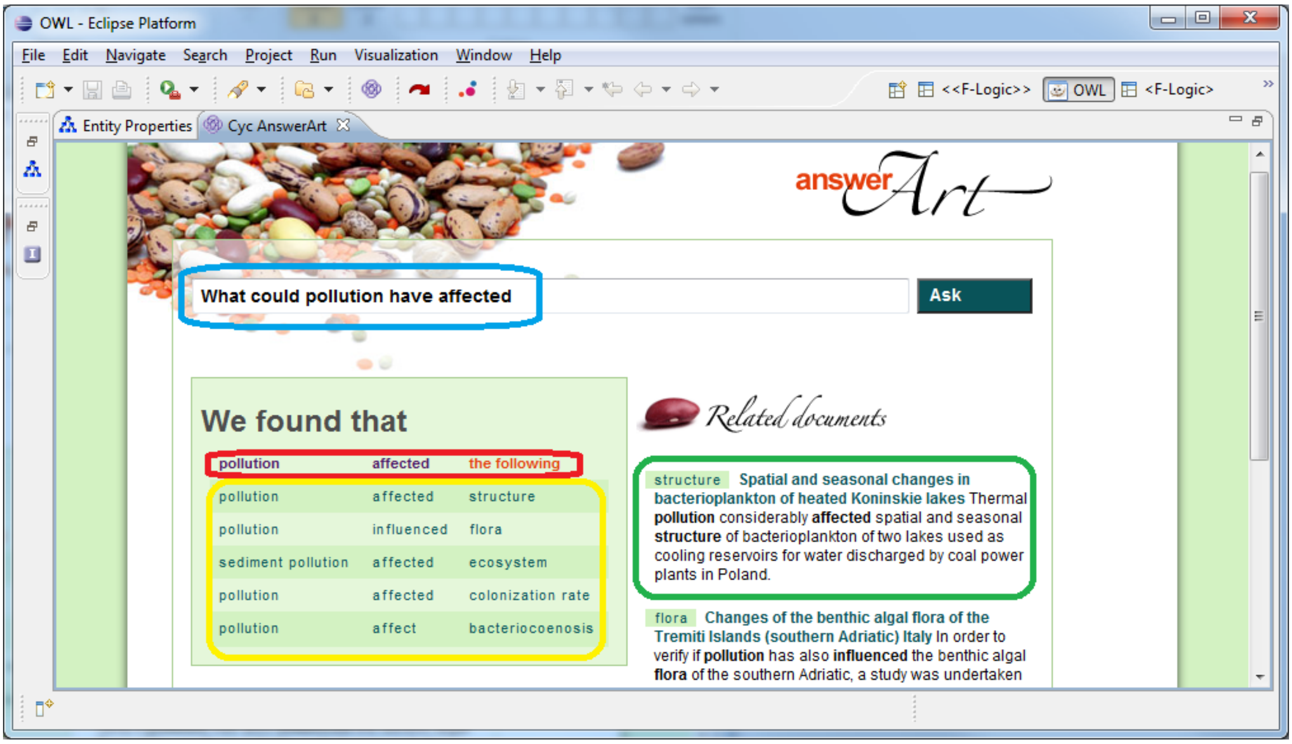
pollution	affected	the following
pollution	affected	structure
pollution	influenced	flora
sediment pollution	affected	ecosystem
pollution	affected	colonization rate
pollution	affect	bacteriocoenosis

Related documents

structure **Spatial and seasonal changes in bacterioplankton of heated Koninskie lakes** Thermal **pollution** considerably **affected** spatial and seasonal **structure** of bacterioplankton of two lakes used as cooling reservoirs for water discharged by coal power plants in Poland.

flora **Changes of the benthic algal flora of the Tremiti Islands (southern Adriatic) Italy** In order to verify if **pollution** has also **influenced** the benthic algal **flora** of the southern Adriatic, a study was undertaken

Figure 5: Returned answers


OWL - Eclipse Platform
File Edit Navigate Search Project Run Visualization Window Help
<<F-Logic>> OWL <F-Logic>
Entity Properties
Cyc AnswerArt
answerArt
Ask
Related documents

What could pollution have affected

We found that

pollution	affected	the following
pollution	affected	structure
pollution	influenced	flora
sediment pollution	affected	ecosystem
pollution	affected	colonization rate
pollution	affect	bacteriocoenosis

structure Spatial and seasonal changes in bacterioplankton of heated Koninskie lakes Thermal **pollution** considerably **affected** spatial and seasonal **structure** of bacterioplankton of two lakes used as cooling reservoirs for water discharged by coal power plants in Poland.

flora Changes of the benthic algal flora of the Tremiti Islands (southern Adriatic) Italy In order to verify if **pollution** has also **influenced** the benthic algal **flora** of the southern Adriatic, a study was undertaken

Figure 6: Detailed description of the plug-in

Here follows the detailed description of the GUI elements (Figure 6):

The blue rectangle marks the input field with the last question asked shown. It is possible to modify or change the question and resubmit it with the "Ask" button to the right of the field.

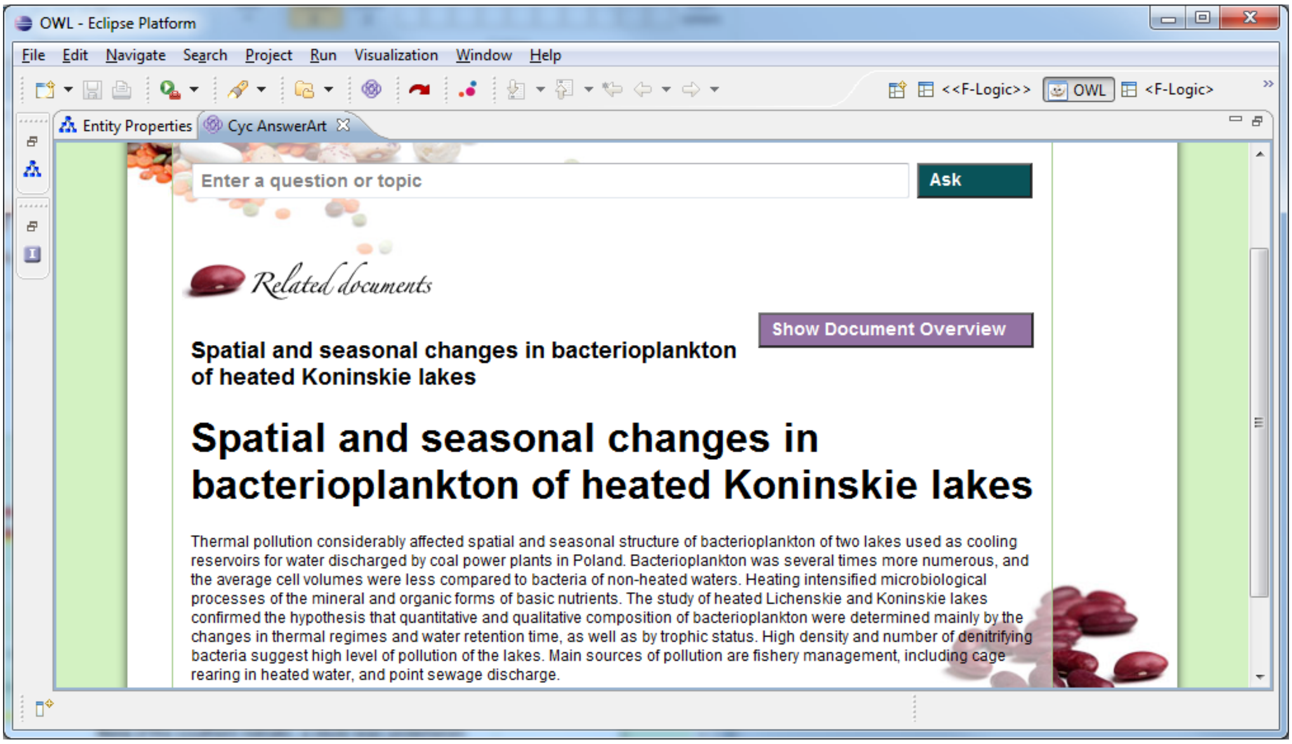
The red rectangle marks the header of the table of results. It visualizes how the natural language question has been transformed into a triples query. The first word designates the subject, the second verb and the last the object. In our example the subject is "pollution" the verb is "affected" and we are searching for all possible objects, which is marked in red colour "the following".

In general it is possible to have missing any single or any two elements from the triple. It is also possible to have all three defined and the question actually checks weather the question is true - or more accurately - if such a claim is found in the document repository. Any defined elements from triples are always printed in black, any missing elements always in red.

The yellow rectangle marks the table of results. Each row is a found triple that satisfies the triple query given in the header. It is however possible, that the specific elements differ. This is due to the semantic triple enhancement (Section 5).

For example, in the third row, the subject is "sediment pollution" found in the ASFA ontology to be a special case of pollution. In the second row the verb is "influenced". This concept was related with the verb "affected" in the Cyc ontology. In the last row the verb is "affect" which is identified with "affected" with basic text mining techniques (stemming, lemmatizing). Any triple element can actually come from the semantically enhanced collection.

To the right are excerpts from the actual documents that the triples were extracted from. The first result is marked with a green rectangle. At the top, the title is shown and clicking leads to the detailed view of the specific document (Figure 7). Bellow is the actual sentence from which the triple was extracted. In our example the sentence for the triple "pollution->affected->structure" is: "Thermal pollution considerably affected spatial and seasonal structure of bacterioplankton ...". The triple elements are marked in bold.


OWL - Eclipse Platform
File Edit Navigate Search Project Run Visualization Window Help
<<F-Logic>> OWL <F-Logic>
Entity Properties
Cyc AnswerArt
Enter a question or topic
Ask
Related documents
Show Document Overview

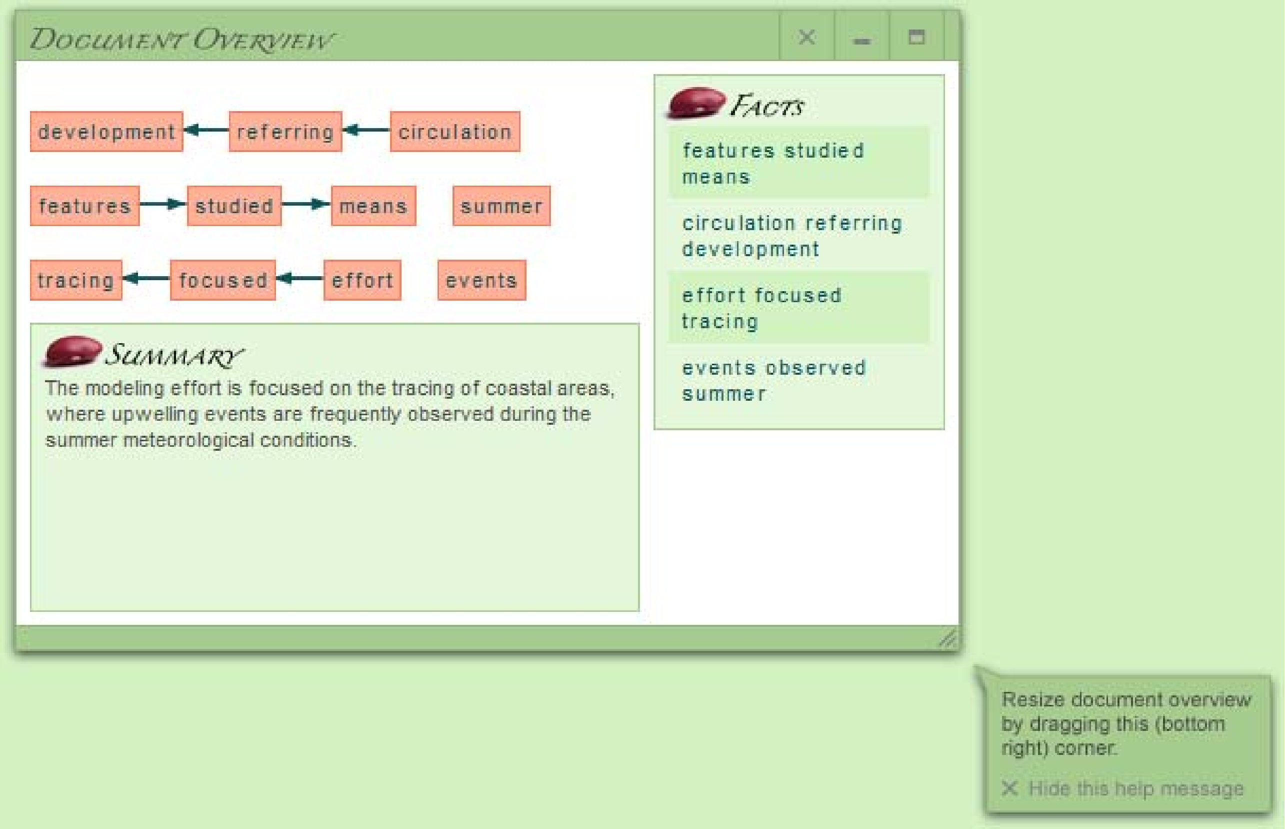
Spatial and seasonal changes in bacterioplankton of heated Koninskie lakes

Spatial and seasonal changes in bacterioplankton of heated Koninskie lakes

Thermal pollution considerably affected spatial and seasonal structure of bacterioplankton of two lakes used as cooling reservoirs for water discharged by coal power plants in Poland. Bacterioplankton was several times more numerous, and the average cell volumes were less compared to bacteria of non-heated waters. Heating intensified microbiological processes of the mineral and organic forms of basic nutrients. The study of heated Lichenskie and Koninskie lakes confirmed the hypothesis that quantitative and qualitative composition of bacterioplankton were determined mainly by the changes in thermal regimes and water retention time, as well as by trophic status. High density and number of denitrifying bacteria suggest high level of pollution of the lakes. Main sources of pollution are fishery management, including cage rearing in heated water, and point sewage discharge.

Figure 7: Document View

Upon clicking on the title of a specific document (Figure 6) a complete document is visualized so it can be studied by the user, to obtain more specific information on the answer. If the user clicks on the "Show Document Overview", a summary view of the document gets visualized (Figure 8, 9).



The modeling effort is focused on the tracing of coastal areas, where upwelling events are frequently observed during the summer meteorological conditions.

Resize document overview by dragging this (bottom right) corner.

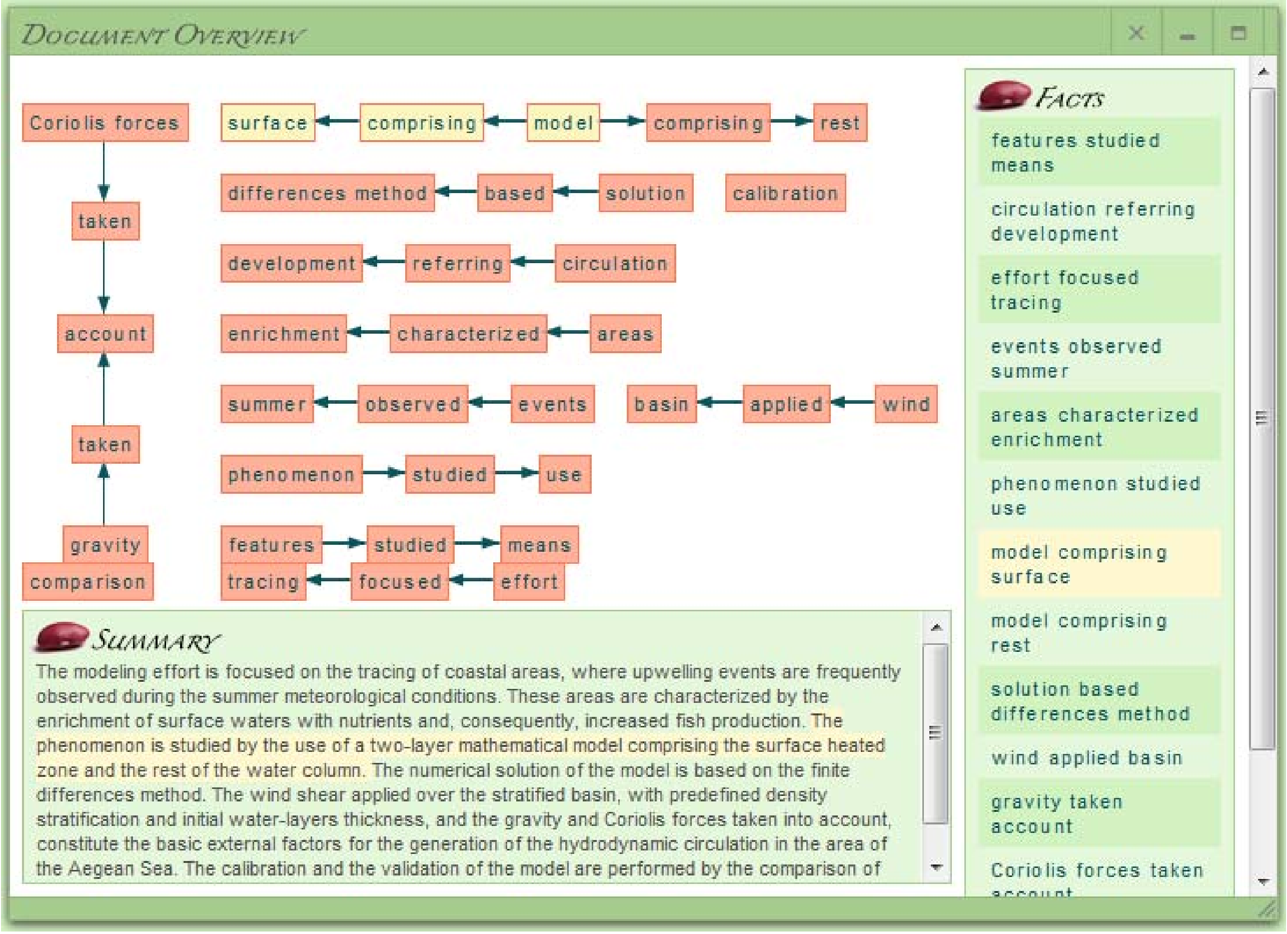
✕ Hide this help message

Figure 8: Summary of the document

In Figure 8 we can see a summary of the document. In the top left corner the most prominent triples are visualized, connected by arrows flowing from subject to verb to object. On the right (Facts) this triples are printed out and at the bottom left this triples are transformed back to natural language.

In our example we are studying document: “Spatial and seasonal changes of heated Koninskie lakes". The calculated summary "The modelling effort is focused on tracing of coastal areas, where..." provides the most important information: the procedure (modelling), location (coastal areas) and time (summer) of the document.

It is also possible to adapt the size of the summary, by dragging the corner of the screen (Figure 8, bottom right). According to the available space the amount of triples and natural language text is visualized. In Figure 9, we can see a bigger summary of the same document.



Summary

The modeling effort is focused on the tracing of coastal areas, where upwelling events are frequently observed during the summer meteorological conditions. These areas are characterized by the enrichment of surface waters with nutrients and, consequently, increased fish production. The phenomenon is studied by the use of a two-layer mathematical model comprising the surface heated zone and the rest of the water column. The numerical solution of the model is based on the finite differences method. The wind shear applied over the stratified basin, with predefined density stratification and initial water-layers thickness, and the gravity and Coriolis forces taken into account, constitute the basic external factors for the generation of the hydrodynamic circulation in the area of the Aegean Sea. The calibration and the validation of the model are performed by the comparison of

Facts

- features studied means
- circulation referring development
- effort focused tracing
- events observed summer
- areas characterized enrichment
- phenomenon studied use
- model comprising surface
- model comprising rest
- solution based differences method
- wind applied basin
- gravity taken account
- Coriolis forces taken account

Figure 9: Extended summary of the document

Triples can also form bigger graph structures. Two or more triples can share an object, in our example "account" seen on the left or a subject ("model" top).

We have also implemented a visualization technique which highlights connected triples and sentences. Upon mouse over any element, sentence or triple the related entities change background colour to yellow. In our example the mouse was pointing over the triple "model comprising surface" in the right "Facts" box. Notice the highlights in graph and the textual Summary.

8. Discussion and future work

The proposed approach to integrating Cyc and AnswerArt technology for contextualized answering of questions provided in English can be adjusted to other natural languages, as all the language specific parts are placed in separate modules.

However, Cyc ontology is in English so the enhancement of the document content should be ensured. This can be done either by translating the document content and questions to English and operating in English or finding some other way to match different natural languages. In case the document content is not translated to English, then the necessary natural language processing to extract triplets should be adjusted for the target natural language.

There is also a lot of room for improvement of the module for transforming natural language question into triples query. Basic improvement could be done in the form of adding a large number of heuristic rules, more sophisticated approach could train on a labelled training set of questions - triple queries and mine patterns in the natural language questions.

The module for triples extraction could be modified to consume different text mining tools (parser, stemmer, lemmatizer) in order to boost the accuracy of triples extraction.

All in all, every module uses state of the art technologies that we have shown work in a complex pipeline to make available question answering on top of document repository with the context of CYC ontology. Some of the modules were already evaluated, while the whole pipeline still needs thorough evaluation experiments, which are part of the future work.

References

[Banko and Etizioni, 2008] Banko, M. and Etzioni, O. 2008. The Tradeoffs Between Open and Traditional Relation Extraction. In Proceedings of the Association for Computational Linguistics, Columbus, Ohio.

[Baxter et al., 2009] David Baxter, Bryan Klimt, Marko Grobelnik, David Schneider, Michael Witbrock, Dunja Mladenić. Capturing Document Semantics for Ontology Generation and Document Summarization. In: Davies, J., Grobelnik M, Mladenić, D (eds.). *Semantic knowledge management : integrating ontology management, knowledge discovery, and human language technology*. Berlin; Heidelberg: Springer, 2009, pp. 141-154.

[Dali et al., 2009] Lorand Dali, Delia Rusu, Blaž Fortuna, Dunja Mladenić, Marko Grobelnik. Question Answering Based on Semantic Graphs. In proceedings of the WWW-2009 Workshop on Semantic Search (SemSearch2009), Madrid, April 2009. <http://km.aifb.uni-karlsruhe.de/ws/semsearch09/semse2009_22.pdf>

[Grobelnik et al., 2007] Marko Grobelnik, Janez Brank, Blaž Fortuna, Igor Mozetič. D3.2.1 Reasoning with contexts – prototype interpreter, NeOn Deliverable 2007

[Grobelnik et al., 2008] Marko Grobelnik, Janez Brank, Blaž Fortuna, Igor Mozetič. Contextualizing Ontologies with OntoLight: A Pragmatic Approach, Informatica Journal. Janauray 2009.

[Lenat, 1995] Lenat D B. Cyc: A Large-Scale Investment in Knowledge Infrastructure. Communications of the ACM 38, no. 11, November 1995

[Steingbach et all, 2000] Steinbach, M., Karypis, G. and Kumar, V. (2000). Acomparison of document clustering techniques. Proc. KDD Workshop on Text Mining. (eds. Grobelnik, M., Mladenić, D. and Milic-Frayling, N.), Boston, MA, USA, 109–110.

[Suárez-Figueroa, 2007] Suárez-Figueroa M.C., Cea G.A., Buil C., Caracciolo C., Dzbor M., Gómez-Pérez A., Herrrero G., Lewen H., Montiel-Ponsoda E., Presutti V. D5.3.1 NeOn Development Process and Ontology Life Cycle, NeOn Deliverable 2007.

[Fagetti E., Privett D.W. & Sears J.R.L., 2009] Aquatic Sciences and Fisheries Thesaurus. Descriptors Used in the Aquatic Sciences and Fisheries Information System, Food and Agriculture Organization of the United Nations, Rome, 2009.

CSA ASFA Database Guide, http://www.csa.com/factsheets/supplements/asfaguide.pdf

ASFA Thesaurus, http://www.csa.com/factsheets/supplements/asfathes.php