KYOTO Annual Report 2009

“Knowledge Yielding Ontologies for Transition-based Organization”

http://www.kyoto-project.eu/

KYOTO makes knowledge sharable between communities of people, across cultures and languages and it makes this knowledge understandable to computers, by assigning meaning to text and giving text to meaning.

1. Introduction

The globalization of markets and communication brings with it a concomitant globalization of world-wide problems and the need for new solutions. Timely examples are global warming, climate change and other environmental issues related to rapid growth and economic developments. Environmental problems can be acute, requiring immediate support and action, relying on information available elsewhere. Knowledge sharing and transfer are also essential for sustainable growth and development on a longer term. In both cases, it is important that distributed information and experience can be re-used on a global scale. The globalization of problems and their solutions requires that information and communication be supported across a wide range of languages and cultures. Such a system should furthermore allow both experts and laymen to access this information in their own language, without recourse to cultural background knowledge.

The goal of KYOTO is a system that allows people in communities to define the meaning of their words and terms in a shared Wiki platform so that it becomes anchored across languages and cultures but also so that a computer can use this knowledge to detect knowledge and facts in text. Whereas the current Wikipedia uses free text to share knowledge, KYOTO will represent this knowledge so that a computer can understand it too. For example, the notion of environmental *footprint* will become defined in the same way in all these languages but also in such a way that the computer knows what information is necessary to calculate a *footprint*.

With these definitions it will be possible to find information on footprints in documents, websites and reports so that users can directly ask the computer for actual information in their environment, e.g. what is the footprint of their town, their region or their company.

The KYOTO system works in 6 steps, as shown in Figure-1:

1. People from a domain specify the locations of diverse and distributed sources of knowledge in different languages. They can do this through a Semantic Wiki environment called Wikyplanet.
2. The text in various languages is captured from the sources and offered to the KYOTO system
3. Term yielding robots (so-called **Tybots**) automatically extract all the important terms and possible semantic relations and relate these to existing semantic networks (so-called **Wordnets**) in each language.
4. The wiki-environment (so-called **Wikyoto**) allows the domain people to maintain the terms and concepts and agree on their meaning within the community and across languages. The meanings are formalized in a domain **ontology** which can be used by computer programs.
5. Knowledge yielding robots (so-called **Kybots**), use the terms and knowledge to detect factual data in the text in various languages.
6. The factual data is indexed and can be accessed by anybody through semantic search, again in various languages, e.g. *facts on CO2 emission in Europe from 2000 to 2009*.

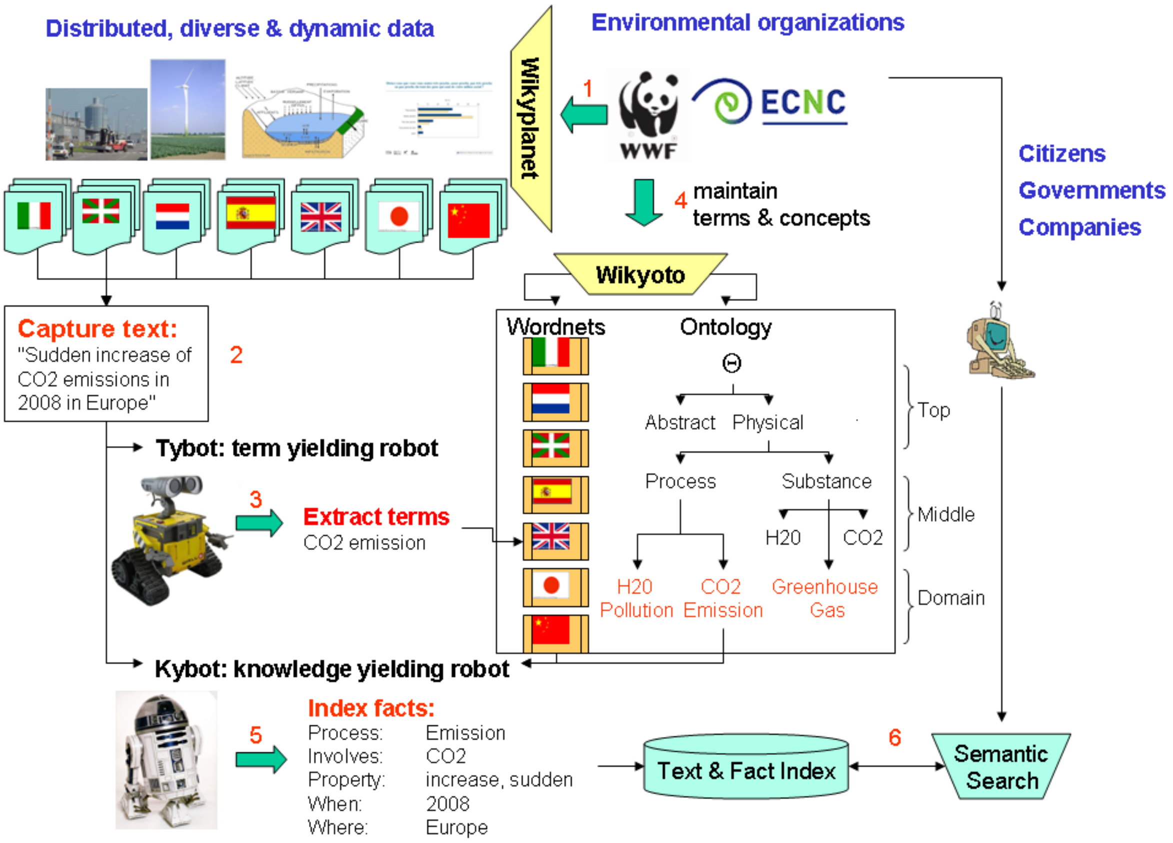


Figure 1: Overall system overview

2. Summary of Activities

Towards the end of the first year of the project (February 2009), we organized a workshop that brought together people from the environment domain and technology experts. The workshop revealed that little technology is used by the environment experts, while the pressure to acquire actual information on specific regions is growing. As a response, we launched a community platform, called Wikyplanet, to organize the community and to stimulate information sharing and dissemination:

http://wikyoto.irion.nl/wikis/wikyoto/index.php/Main_Page.

Wikyplanet will be the central platform for disseminating KYOTO to a wider audience. As a user case, we have set up Wiki pages for estuaries in different countries that provide the data for modeling concepts and knowledge in KYOTO. Some of these cases were used to define the terms and concepts and the knowledge that matter for the environment.

Furthermore, we completed the cycle of knowledge for the first databases and developed a first version of a semantic search system through which people from the domain can get access to this knowledge. The search system can be accessed through:

http://index.irion.nl/kyoto-1.0-SNAPSHOT/

The semantic search shows that structured knowledge can be retrieved and displayed in more useful ways than regular Google-like searches. Searches are made cross-lingual through the multilingual wordnet architecture.

A third major achievement of the project is the integration of large volumes of background knowledge with the knowledge that KYOTO extracts from the text. We adapted the design of the knowledge repository and developed software to automatically align and integrate background vocabulary into the system. This was applied to a database with 2.1 million species. This also redefined the work for the domain experts to model their terms and concepts (step 4 in the above overview). Instead of handling large volumes of background concepts, the domain experts can focus on smaller selections of terms of their interest. The crucial aspect is that these need to be related to the processes and properties that matter. For that purpose, we developed an extension of the central ontology for these processes and properties and adapted the Wikyoto editing environment so that these mappings can be created in an easy way.

Finally, we have created a pipeline architecture for the various KYOTO modules and will distribute the first release of the integrated system by the end of the second year of the project. This will allow developers outside the project to use and test the system and possibly to extend it to other languages.

The third and final year of the project will be devoted to improving and fine tuning the complete KYOTO system through several evaluation and development rounds. Furthermore, we will develop more demonstration systems that show the exploitation possibilities of the system. At the end of the project, we will organize a workshop for a wider audience to present the final project results.

3. User involvement and requirements

In February 2-3, 2009, we organized a two day workshop "Environmental Knowledge Transition and Exchange" in Amsterdam bringing together specialists in the environment domain and technology developer. We invited 4 keynote speakers from different fields. The workshop was structured to deliver the technical partners with an understanding of the problems and wishes in regard to information gathering within the environmental domain on the one hand and to explain the possibilities of information technology to people in the environmental domain on the other hand. This resulted in a better understanding of the needs of the environmental domain.

The workshop provided a significant insight and an increased level of awareness in relation to the clear "technology gap" currently suffered by the environment community. Compared to other areas, such as for example the bio-technology and medicine area, the environment community is apparently using very little information technology in general and, where it does, makes use of the basic technology in a non-optimal way. The workshop also made it clear that a change of attitude and mentality is needed. The people in the environment domain should become more aware of the limitations of the current methodologies used and they should be more willing to share resources and information in the community to the greater good.

A third point raised is that the pressure to process massive amounts of rapidly changing information is enormous. The users need to be able to instantly acquire actual and comprehensive information from a diverse range of sources to be able to fulfill their tasks. This information often needs to be acquired for very local and unique situations and cannot be derived from general sources. Using the standard web technology therefore leads to a major risk that conclusions are based on shallow information searches, (i.e. not going beyond the first result page of Google – a statistic confirmed by the users).

Another major conclusion is that the users have a very urgent need to organize themselves as an Internet community, preferably through a community platform.

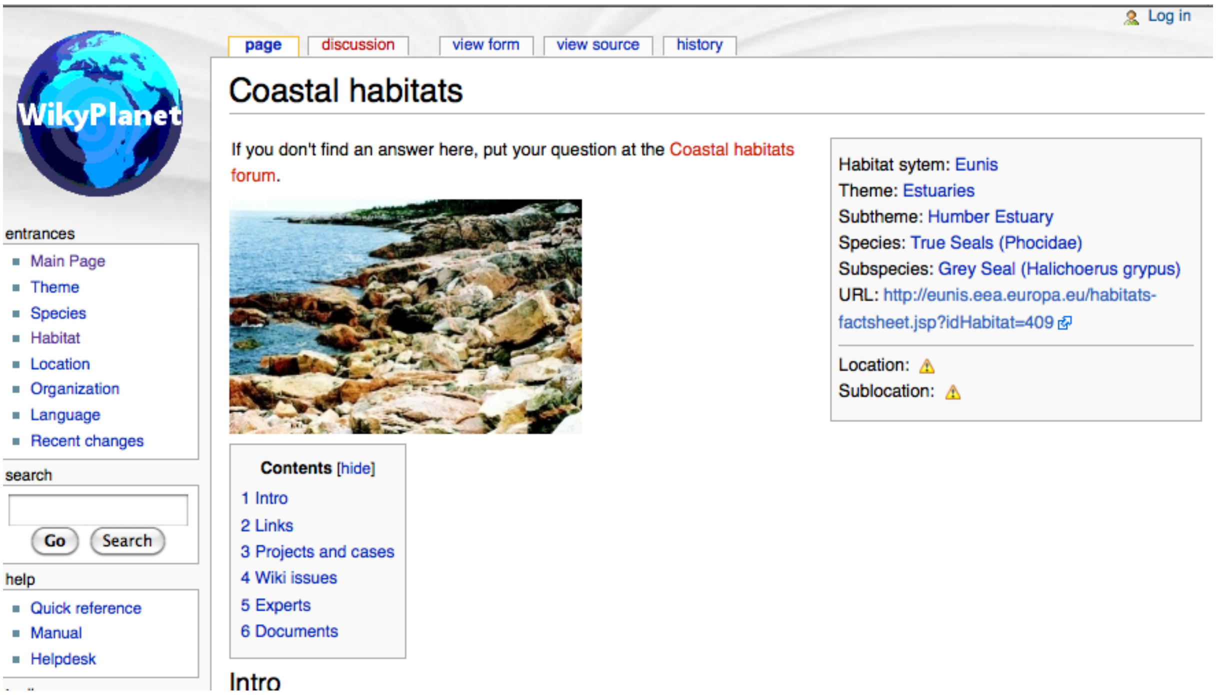
Given the enormous information growth and the time and delivery pressure on processing this information, it becomes crucial that information and knowledge transfer is efficient and correct. Politicians depend on the correctness of reports from environmentalist that again depend on factual data provided by the scientific community. Wrongly acquired information can be fatal to decision processes. It is clear that the KYOTO system can contribute to providing specialists at each of these levels with better, more comprehensive and more actual information and knowledge.

The workshop resulted into two immediate actions within the project:

1. We implemented a Semantic Wiki environment: http://wikyoto.irion.nl/wikis/wikyoto/index.php/Main_Page, that can directly be used as a Semantic Web2.0 environment for the community to share knowledge and nformation;
2. We collected specific sources and data related to estuaries in different areas in the world for which so-called mind maps were constructed, reflecting the concepts and data that are of interest to the users;

The semantic wiki environment called Wikyplanet distinguishes four categories of information: theme, habitat, species and location. Each of these categories is organized hierarchically into more specific pages. Domain specialists can add information to any category and create cross links to any other categories.

In the next screen dump, you can see the page for Coastal habitats. It is a subpage of Habitats and is connected to the themes Estuaries and Humber Estuary, and to the species True Seals. The page has further slots to specify descriptions, links, projects and cases. Users can also directly post questions to the forum from this page.



Coastal habitats

If you don't find an answer here, put your question at the Coastal habitats forum.

Habitat sytem: Eunis
Theme: Estuaries
Subtheme: Humber Estuary
Species: True Seals (Phocidae)
Subspecies: Grey Seal (Halichoerus grypus)
URL: http://eunis.eea.europa.eu/habitats-factsheet.jsp?idHabitat=409
Location: ⚠
Sublocation: ⚠

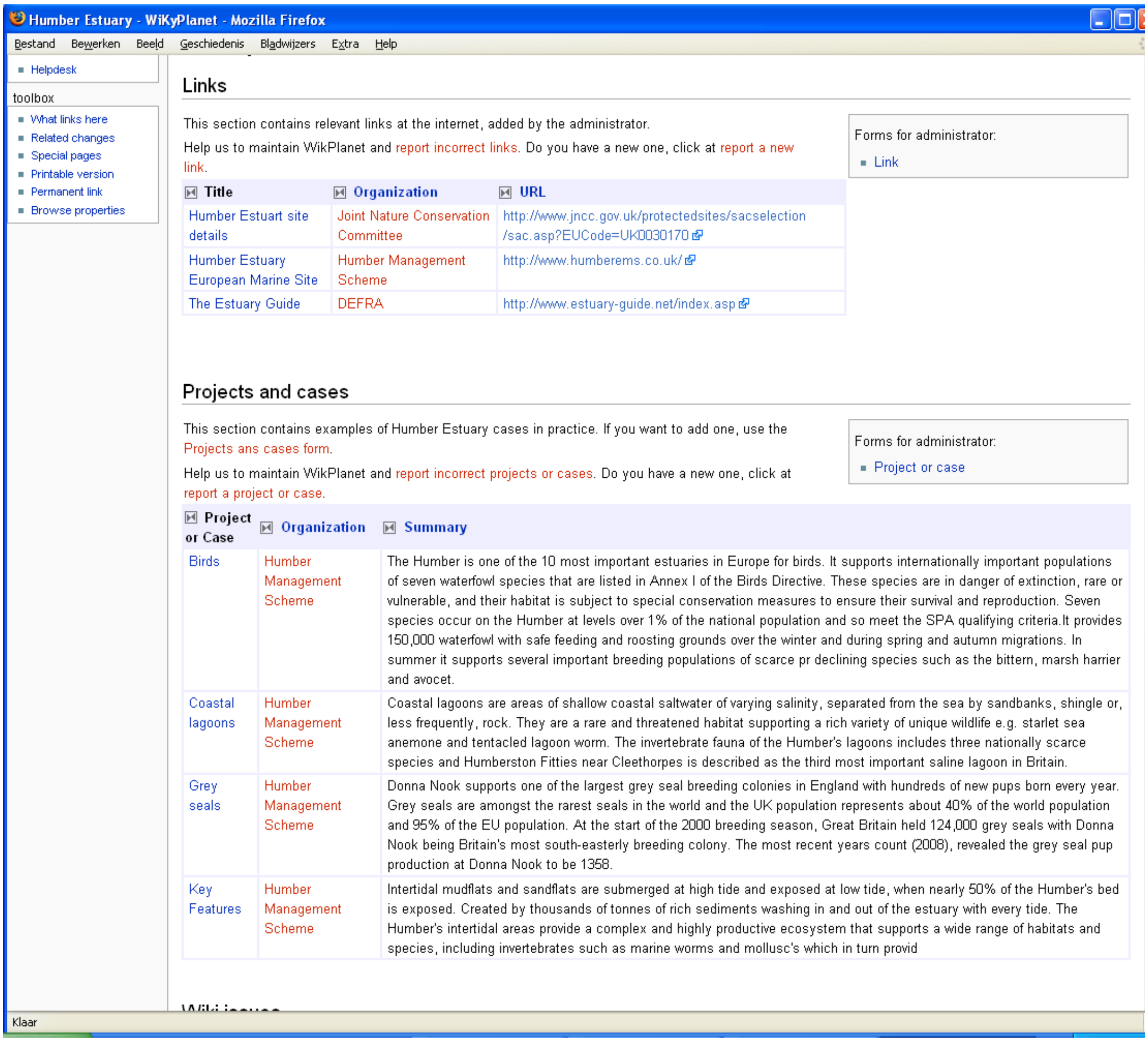
Contents [hide]
1 Intro
2 Links
3 Projects and cases
4 Wiki issues
5 Experts
6 Documents

Intro

Figure 2: Wikyplanet page for Coastal habitats

The following screen dumps show the linked pages of Grey Seals and the Humber Estuary, which has further information and cross-links. They include specific descriptions, links to sources, websites, projects and case descriptions. Each landing page can thus be embedded with cross-links to other types of information and more specific or more general pages of the same type. Separate hierarchies of each category can be created independently and cross-links can be created at any point. The current platform is available in English but other language versions of the system will be released soon.

The Wikyplanet system allows users to gather and share information in an easy to use manner. Wikyplanet will be used as the basis for launching the KYOTO system itself. Data provided on the wiki pages will be processed by KYOTO so that the entailed knowledge can be mined and made available to the community. Wikyplanet is currently launched within the project user organizations ECNC and WWF. In the next phase of the project, we will distribute it to a wider audience in Europe and the rest of the world.



Links

This section contains relevant links at the internet, added by the administrator.

Help us to maintain WikPlanet and report incorrect links. Do you have a new one, click at report a new link.

Forms for administrator:

- Link

Title	Organization	URL
Humber Estuart site details	Joint Nature Conservation Committee	http://www.jncc.gov.uk/protectedsites/sacselection/sac.asp?EUCode=UK0030170
Humber Estuary European Marine Site	Humber Management Scheme	http://www.humberems.co.uk/
The Estuary Guide	DEFRA	http://www.estuary-guide.net/index.asp

Projects and cases

This section contains examples of Humber Estuary cases in practice. If you want to add one, use the Projects ans cases form.

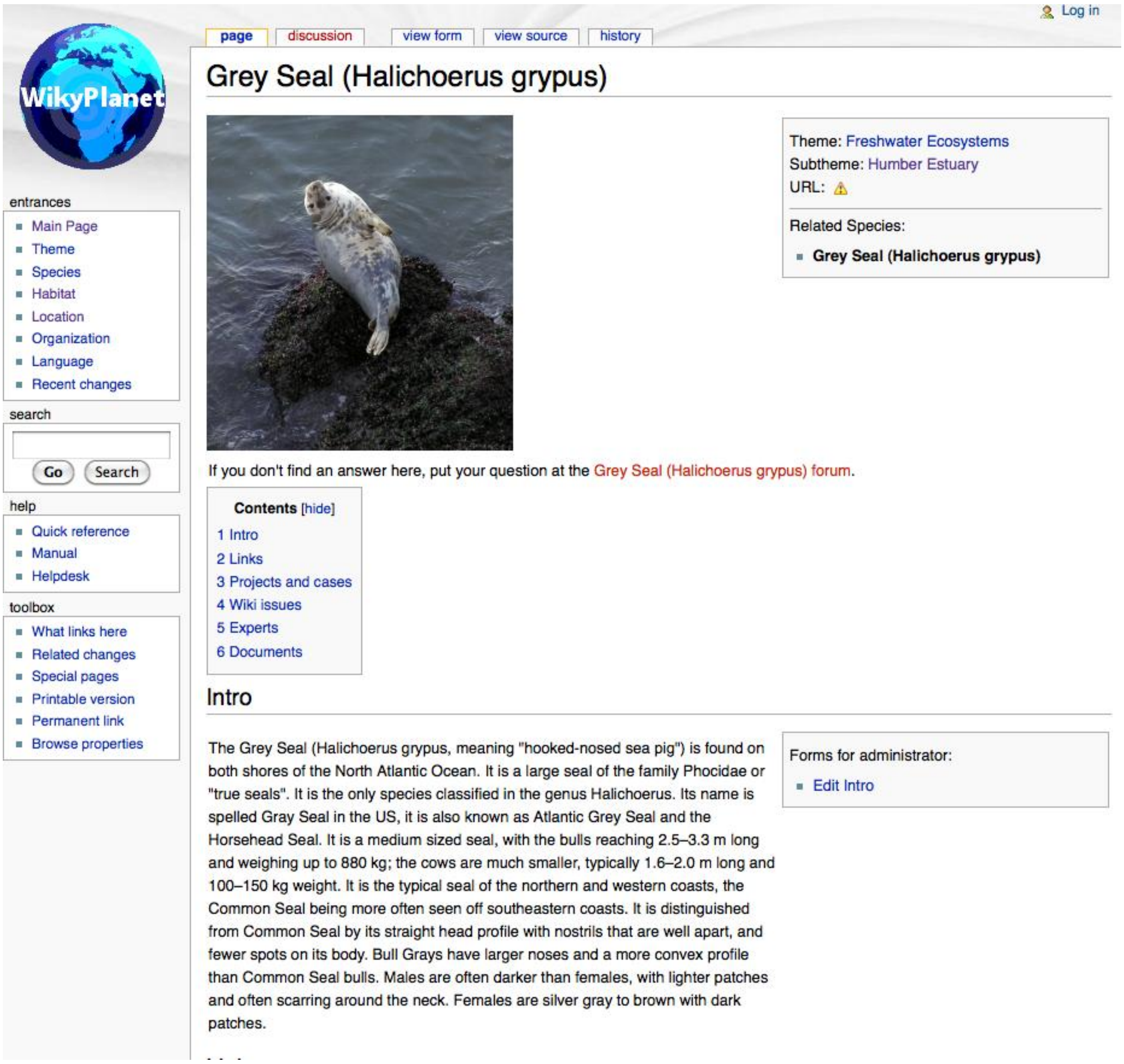
Help us to maintain WikPlanet and report incorrect projects or cases. Do you have a new one, click at report a project or case.

Forms for administrator:

- Project or case

Project or Case	Organization	Summary
Birds	Humber Management Scheme	The Humber is one of the 10 most important estuaries in Europe for birds. It supports internationally important populations of seven waterfowl species that are listed in Annex I of the Birds Directive. These species are in danger of extinction, rare or vulnerable, and their habitat is subject to special conservation measures to ensure their survival and reproduction. Seven species occur on the Humber at levels over 1% of the national population and so meet the SPA qualifying criteria.It provides 150,000 waterfowl with safe feeding and roosting grounds over the winter and during spring and autumn migrations. In summer it supports several important breeding populations of scarce pr declining species such as the bittern, marsh harrier and avocet.
Coastal lagoons	Humber Management Scheme	Coastal lagoons are areas of shallow coastal saltwater of varying salinity, separated from the sea by sandbanks, shingle or, less frequently, rock. They are a rare and threatened habitat supporting a rich variety of unique wildlife e.g. starlet sea anemone and tentacled lagoon worm. The invertebrate fauna of the Humber's lagoons includes three nationally scarce species and Humberston Fitties near Cleethorpes is described as the third most important saline lagoon in Britain.
Grey seals	Humber Management Scheme	Donna Nook supports one of the largest grey seal breeding colonies in England with hundreds of new pups born every year. Grey seals are amongst the rarest seals in the world and the UK population represents about 40% of the world population and 95% of the EU population. At the start of the 2000 breeding season, Great Britain held 124,000 grey seals with Donna Nook being Britain's most south-easterly breeding colony. The most recent years count (2008), revealed the grey seal pup production at Donna Nook to be 1358.
Key Features	Humber Management Scheme	Intertidal mudflats and sandflats are submerged at high tide and exposed at low tide, when nearly 50% of the Humber's bed is exposed. Created by thousands of tonnes of rich sediments washing in and out of the estuary with every tide. The Humber's intertidal areas provide a complex and highly productive ecosystem that supports a wide range of habitats and species, including invertebrates such as marine worms and mollusc's which in turn provid

Figure 3: Wikyplanet linked page with further information



Grey Seal (Halichoerus grypus)

Theme: Freshwater Ecosystems
Subtheme: Humber Estuary
URL: ⚠

Related Species:

- Grey Seal (Halichoerus grypus)

If you don't find an answer here, put your question at the Grey Seal (Halichoerus grypus) forum.

Contents [hide]
1 Intro
2 Links
3 Projects and cases
4 Wiki issues
5 Experts
6 Documents

Intro

The Grey Seal (Halichoerus grypus, meaning "hooked-nosed sea pig") is found on both shores of the North Atlantic Ocean. It is a large seal of the family Phocidae or "true seals". It is the only species classified in the genus Halichoerus. Its name is spelled Gray Seal in the US, it is also known as Atlantic Grey Seal and the Horsehead Seal. It is a medium sized seal, with the bulls reaching 2.5–3.3 m long and weighing up to 880 kg; the cows are much smaller, typically 1.6–2.0 m long and 100–150 kg weight. It is the typical seal of the northern and western coasts, the Common Seal being more often seen off southeastern coasts. It is distinguished from Common Seal by its straight head profile with nostrils that are well apart, and fewer spots on its body. Bull Grays have larger noses and a more convex profile than Common Seal bulls. Males are often darker than females, with lighter patches and often scarring around the neck. Females are silver gray to brown with dark patches.

Forms for administrator:

- Edit Intro

Figure 4: Wikyplanet linked page with further information

Within Wikyplanet, we started to collect information on specific estuaries in the world. This has resulted in data and sources for the following estuaries:

- English: Humber Estuary in the UK, and the Chesapeak Bay in the US. Further background data on migration birds, sedimentation and habitat loss.
- Spanish: Delta del Ebro
- Basque: Urdaibai
- Italian: Poo Estuary
- Dutch: Westerschelde

We built specific databases for each language which will be used for the first evaluation of the complete system. ECNC and WWF created a mind map for the Humber Estuary case using some of the documents as a basis. The mind maps represent the relevant terms and concepts that link to important processes and features for the region. Figure 5 shows a fragment of such a mind map.

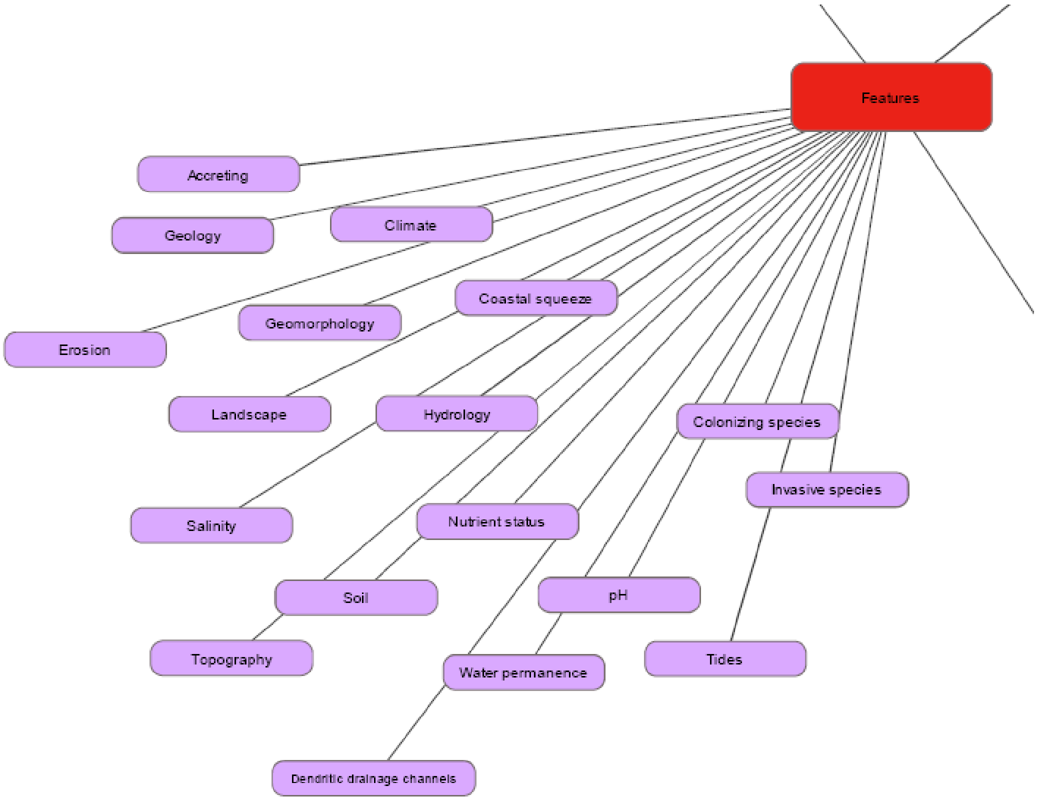


Figure 5: Snaphot view of mind map for the Humber Estuary

The mind maps will form the basis for evaluating the KYOTO system. KYOTO should be able to automatically extract similar information and knowledge for the different estuary databases. This evaluation will be carried out before the end of the second year of the project.

4. System and knowledge architecture

In the first year of the project, we developed first versions of all the modules, which have been applied and tested for 7 languages in the KYOTO project. In the second year, we have further integrated these modules in a single pipeline architecture that is hosted on a central server. This is a first step towards a fully integrated system. The first release of the integrated system is expected by the end of the second year of the project.

The pipeline architecture is organized around a central database for the documents that contain the important information. The architecture is open and flexible so that developers can include any natural language processing tool that they want to include. Likewise, the system can easily be extended to other languages as well. The key of the system is the representation of the processed text in the KYOTO Annotation Format (KAF). This is an open format for representing the structure and meaning of text through separate annotation layers. This representation is compatible with LAF and GRAF which are data models proposed in ISO working groups.

Once textual sources are represented in this format, a series of KYOTO modules can be applied that operate in the same manner across all the languages:

- The meanings of the individual words is detected through word-sense-disambiguation software that uses the wordnet structure as a basis. Likewise, the system distinguishes “plants” as species from “plants” as factories in text.

- Named entities are detected using date recognition and the GeoNames database. This makes it possible to connect knowledge to periods and regions and to show timelines and position regions on e.g. Google-maps.
- Language-neutral ontological classes are assigned to the text using the named entities and the word meanings that are detected. This makes it possible to search for patterns regardless of the language in which the knowledge is expressed and to relate knowledge extracted from text in different languages.

This process has been implemented and tested for all the KYOTO languages. The representation of the text reveals the linguistic structures and semantic annotations that are the input for the major KYOTO cycles:

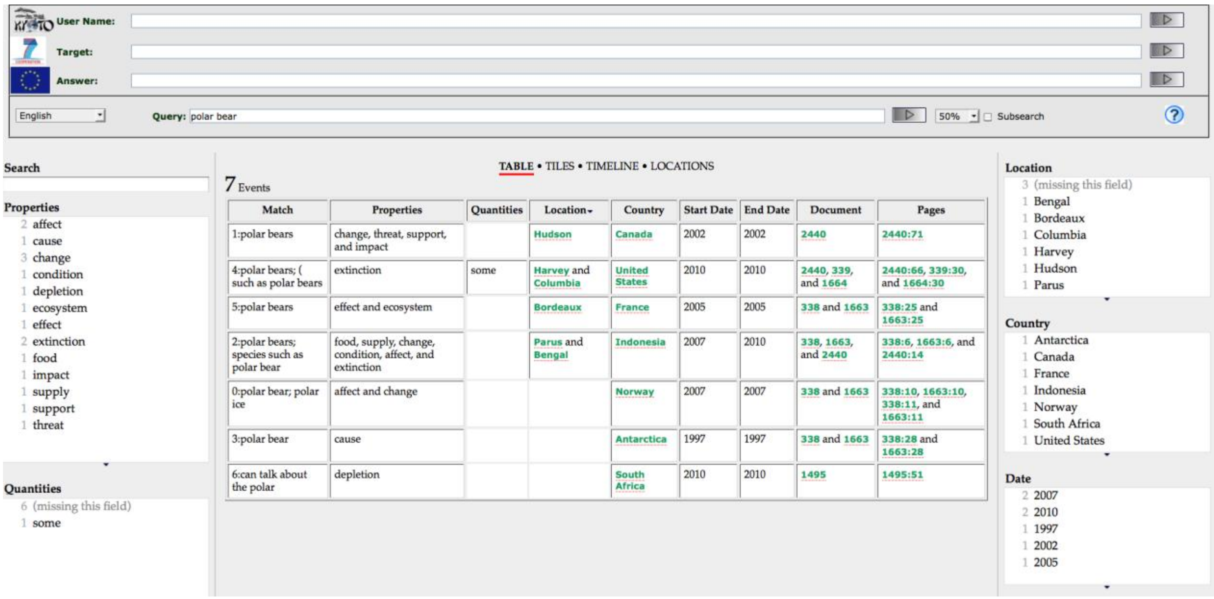
1. Terms and relations between the terms are learned from the representations are automatically linked to wordnet databases and the central ontology. This is done by the Term Yielding Robot (Tybot) in the same way for all the languages. The mined terms and concepts can be viewed and edited by the specialists in the domain, using the Wikyoto platform. This is a user-friendly Wiki platform that does not require specialized knowledge on knowledge engineering.
2. Relations and facts are detected in the text and are represented in a generic format. This is done by the Knowledge Yielding Robot (Kybot) that uses knowledge profiles. The knowledge profiles are layered patterns of linguistic structures and conceptual patterns that reflect structures in the annotated text. Profiles can be language neutral or language specific. The conceptual patterns are taken from the terms and concepts selected by the domain specialists in the first cycle.

We have carried out the complete cycle for a number of databases in the project for which the users provided the sources. Knowledge from sources in different language on similar topics has been mined in a uniform and compatible way.

4. Semantic search

The knowledge that is mined from the documents can be accessed by non-experts and experts in the domain through a semantic search service. The search service takes ordinary queries in natural language and collects all knowledge that relates to it. The main difference from a Google-like interface is that the results are organized as semantic data. The search starts from the assumption that knowledge from the environment is related to a specific region and time frame. Even if queries do not make the time frame and region explicit, we still interpret any properties and relations relative to regions and periods in which they can take place through the information that is detected in the text.

A query for “polar bears” will this not just result in documents and pages where “polar bears” are referred to in the text, but the system collects all the environmental properties and processes in which polar bears are involved and distribute these over different periods and regions that apply to each statement. Figure 6 shows the results in a table representation. Surrounding the table are properties related to polar bears that are extracted from the text, locations and countries that were detected and the relevant periods. The user can select any of these items or combinations of items to filter the related results in the table. Each row in the table then represents a result limited to a defined period and region. In this setting, all results within the range of five years and a single country are represented together. The second row for example groups matches found in two cities in the US, taken from 3 pages belonging to 3 different documents. The relevant property is extinction.



User Name:

Target:

Answer:

English

Query: polar bear

50% □ Subsearch

Search

Properties
- 2 affect
- 1 cause
- 3 change
- 1 condition
- 1 depletion
- 1 ecosystem
- 1 effect
- 2 extinction
- 1 food
- 1 impact
- 1 supply
- 1 support
- 1 threat

Quantities
- 6 (missing this field)
- 1 some

TABLE • TILES • TIMELINE • LOCATIONS

7 Events

Match	Properties	Quantities	Location	Country	Start Date	End Date	Document	Pages
1:polar bears	change, threat, support, and impact		Hudson	Canada	2002	2002	2440	2440:71
4:polar bears; (such as polar bears	extinction	some	Harvey and Columbia	United States	2010	2010	2440, 339, and 1664	2440:66, 339:30, and 1664:30
5:polar bears	effect and ecosystem		Bordeaux	France	2005	2005	338 and 1663	338:25 and 1663:25
2:polar bears; species such as polar bear	food, supply, change, condition, affect, and extinction		Parus and Bengal	Indonesia	2007	2010	338, 1663, and 2440	338:6, 1663:6, and 2440:14
0:polar bear; polar ice	affect and change			Norway	2007	2007	338 and 1663	338:10, 1663:10, 338:11, and 1663:11
3:polar bear	cause			Antarctica	1997	1997	338 and 1663	338:28 and 1663:28
6:can talk about the polar	depletion			South Africa	2010	2010	1495	1495:51

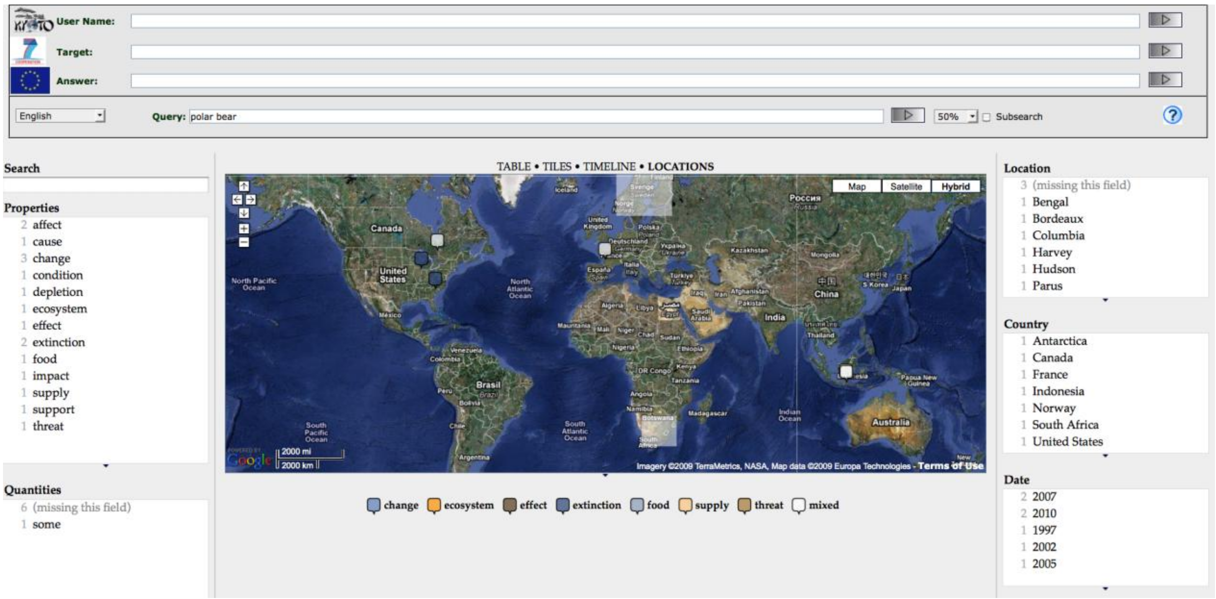
Location
- 3 (missing this field)
- 1 Bengal
- 1 Bordeaux
- 1 Columbia
- 1 Harvey
- 1 Hudson
- 1 Parus

Country
- 1 Antarctica
- 1 Canada
- 1 France
- 1 Indonesia
- 1 Norway
- 1 South Africa
- 1 United States

Date
- 2 2007
- 2 2010
- 1 1997
- 1 2002
- 1 2005

Figure 6: Semantic search results for "polar bear" in the English database

The same results can also be displayed on Google maps and on a time-line as is shown in the next screen dumps. In both cases, the properties related to the polar bear matches are reflected by different colors, as explained in the legend below the display. Each depicted result can be clicked to obtain more details or go to the original document. In the case of the time-line interface, the user can scroll along the time axis and look for temporal relations across properties related to the query.



User Name:

Target:

Answer:

English

Query: polar bear

50% □ Subsearch

Search

Properties
- 2 affect
- 1 cause
- 3 change
- 1 condition
- 1 depletion
- 1 ecosystem
- 1 effect
- 2 extinction
- 1 food
- 1 impact
- 1 supply
- 1 support
- 1 threat

Quantities
- 6 (missing this field)
- 1 some

TABLE • TILES • TIMELINE • LOCATIONS

change ecosystem effect extinction food supply threat mixed

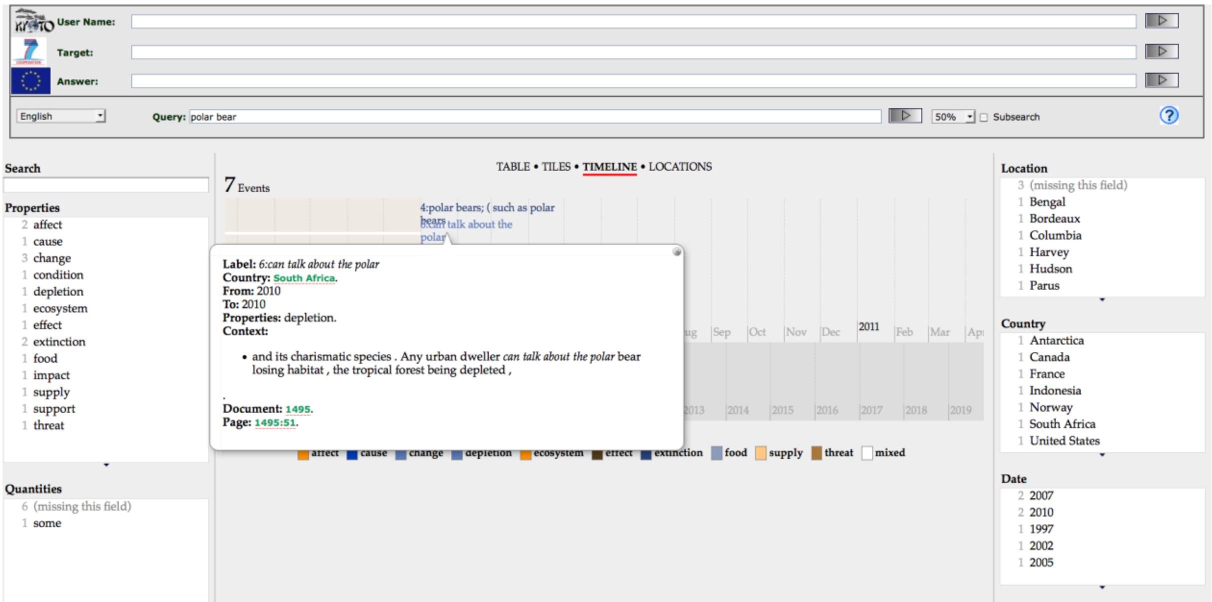
Location
- 3 (missing this field)
- 1 Bengal
- 1 Bordeaux
- 1 Columbia
- 1 Harvey
- 1 Hudson
- 1 Parus

Country
- 1 Antarctica
- 1 Canada
- 1 France
- 1 Indonesia
- 1 Norway
- 1 South Africa
- 1 United States

Date
- 2 2007
- 2 2010
- 1 1997
- 1 2002
- 1 2005

Figure 7: Google map display for semantic search results

At each point in the interface, the users can access the source documents underlying the results. For these documents they can indicate whether it is useful or not. A similar feedback mechanism has been implemented for the baseline search system on the same data. The actions of the user are logged and are used to determine the efficiency of the system for finding relevant information. The evaluation of the system will be carried out before the end of the second year.



User Name:

Target:

Answer:

English

Query: polar bear

50% ☐ Subsearch

Search

Properties
2 affect
1 cause
3 change
1 condition
1 depletion
1 ecosystem
1 effect
2 extinction
1 food
1 impact
1 supply
1 support
1 threat

Quantities
6 (missing this field)
1 some

TABLE • TILES • TIMELINE • LOCATIONS

7 Events

Label: *6:can talk about the polar*
Country: South Africa.
From: 2010
To: 2010
Properties: depletion.
Context:

- and its charismatic species . Any urban dweller *can talk about the polar* bear losing habitat , the tropical forest being depleted ,

.
Document: 1495.
Page: 1495:51.

Location
3 (missing this field)
1 Bengal
1 Bordeaux
1 Columbia
1 Harvey
1 Hudson
1 Parus

Country
1 Antarctica
1 Canada
1 France
1 Indonesia
1 Norway
1 South Africa
1 United States

Date
2 2007
2 2010
1 1997
1 2002
1 2005

Figure 8: Timeline display for semantic search results

5. Domain extension and knowledge integration

The modeling of the important terms and concepts in the domain is an important step towards to the disclosure of the knowledge of a domain. The Wikyoto platform is the module that helps the domain specialists to select and connect the important terms in the their domain. These terms are not only acquired from the documents that are provided by the user. In many cases, communities already have large quantities of (semi-)structured vocabularies and thesauri. For example in the case of the environment domain, the following knowledge repositories are relevant:

- Wikipedia: by September 2009 it has more than 3 million articles in English and large volumes in other languages: http://www.wikipedia.org/.
- DBPedia: by September 2009 it has 2.6 million things, including at least 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies. The knowledge base consists of 274 million pieces of information (RDF triples): http://dbpedia.org/About.
- GeoNames: by September 2009 it has eight million geographical names and consists of 6.5 million unique features whereof 2.2 million populated places and 1.8 million alternate names. All features are categorized into one out of nine feature classes and further subcategorized into one out of 645 feature codes: http://www.geonames.org/about.html.
- The Species 2000 database with 2.1 million species, having taxonomic relations and labels in many different languages: http://www.sp2000.org/.

- A term databases with about 500,000 terms per 1,000 documents in each of the 7 languages.
- Generic wordnets in each language ranging from 50,000 to 120,000 synsets
- Existing ontologies such as the EuroWordNet top-ontology (Vossen 1998), SUMO (Niles and Pease 2002) and DOLCE (Masolo et al 2003) with hundreds up to thousands of formally defined concepts.

Modeling the terms and concepts for a domain for multiple languages is thus a huge integration task involving millions of concepts and relations. To help the modeling, we extended the KYOTO knowledge platform with the option to automatically load and access large background repositories. The repositories need to be made available in SKOS format, which is a standardized W3C format for representing vocabularies and thesauri that cannot be formally defined in RDF of OWL. Such vocabularies are automatically aligned with generic wordnets for languages. This was done for the 2.1 million species in the Species 2000 database. The resulting repository of knowledge thus consists of 3 layers as shown in Figure 9. In this architecture, millions of species are defined in the background vocabulary, which remains in a separate database that is aligned to a generic wordnet database for a language. The terms in the term database are also aligned to the same language wordnet. Each of these databases has its own hierarchical structure. Since we have the same resources and systems for all 7 KYOTO languages, all textual terms in these languages are anchored to the same ontological distinctions.

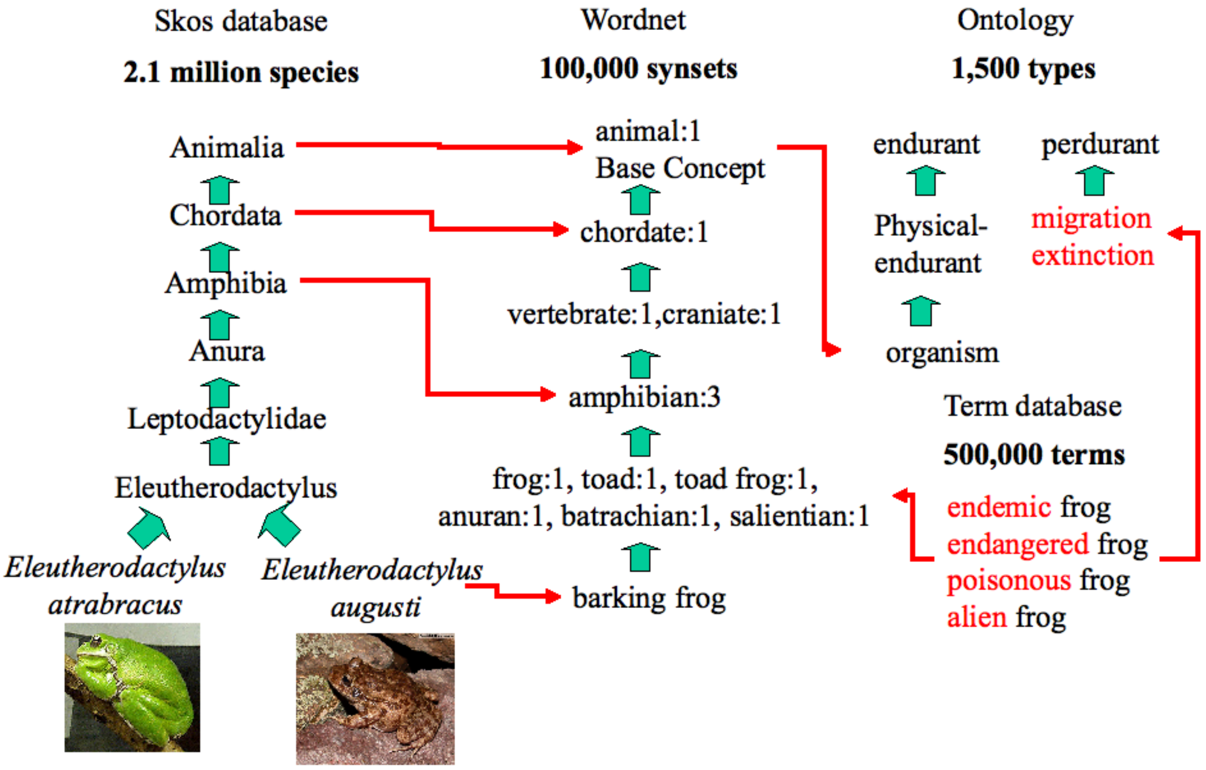


Figure 9: Division of knowledge over 3 layers

For mining facts in a text, users can create a conceptual pattern defined in terms of the language neutral ontology, e.g. “Organism-Invading-CoastalArea”. KYOTO will then first scan the text for terms and match each term with the vocabularies in SKOS or in the term database. If a term is found, we traverse the internal relations in each database to find the first match with the generic wordnet of the language. The wordnet hierarchy is then followed until we find a match with the ontological class. KYOTO then verifies if that class is compatible with the pattern. In the case of a match, we will output the semantic relation to the database of facts with pointers to the sources. Likewise, we can extract facts from text in different languages using the same pattern.

Currently, we developed an extension of the DOLCE ontology that represents the processes and states that are relevant for the domain cases (esp. the estuary databases) and a basic ontology for the type of objects in the area. Verbs and adjectives in the different languages, as well as role denoting terms such as “endangered species”, “invasive species” are mapped to these processes. The ontology is likewise supporting the detection of processes and states that matter to the domain.

6. User Involvement, Promotion and Awareness

Two environmental organizations take an active role in the project: ECNC and WWF. These organizations have however many contacts with other organizations in Europe and across the world. They will be able to organize a wide-scale validation and awareness of the system in their community, beyond the scope of the project. The Wikyplanet system described above will be the primary platform for organizing the wider user community.

The consortium also is represented by the Global WordNet Association (GWA: http://www.globalwordnet.org), which is a non-profit organization that stimulates the development and linking of wordnets for as many languages as possible in the world. Through GWA, the KYOTO system can gain global scale and become a platform for anchoring languages and knowledge mining across the world. We proposed an extension to the LMF ISO standard for representing wordnets: WN-LMF. The wordnets for 7 languages in this format are available in a Global Wordnet Grid architecture, making it easy for other language groups to join the KYOTO framework. These wordnets are all linked to the English WordNet (http://wordnet/princeton.edu) and to a series of language neutral ontologies.

We are cooperating closely with a series of other projects as well. The ontological work is coordinated in close cooperation with the developers of major upper ontologies, e.g. DOLCE. This guarantees a solid foundation of the knowledge representation in the project that is ready for the future and not restricted to an ad-hoc project for a single domain. As for standardization of the system and knowledge repositories, we collaborate closely with the projects Clarin (http://www.clarin.eu/), Flarenet (project ECP-2007-LANG-617001) and Lirics (project 22236 – LIRICS, http://lirics.loria.fr/index.html). In the case of the latter, we have adopted the ISO proposals and recommendations for lexical resources (LMF) and terminology representation (TMF). Furthermore, the KYOTO Annotation Format (KAF) has been based on standards for linguistic annotation as proposed in ISO working groups, such as MAF, SYNAF, SEMAF, LAF and GRAF. KAF is an open representation for linguistically processed text that represents the starting point for all further processes in KYOTO. New languages can be added to the system by converting the output of their linguistic processing to the KAF format and providing a wordnet in WN-LMF that is linked to the English Wordnet.

The results of KYOTO are already used in a number of follow-up projects. The DutchSemCor project (funded by NWO) will use the word-sense-disambiguation technology of KYOTO for building a semantically tagged corpus of Dutch. The Semantics of History project (funded by the VU University Amsterdam through the Camera institute) uses KYOTO to develop the terminology and ontology for an historical archive. KYOTO is also participating in a new FP7 project proposal for the development of a platform for clinical cancer treatment in the medical domain. The Euskal Herriko Unibertsitatea in San Sebastian received funding for the project KNOW2: Language understanding technologies for multilingual domain-oriented information access. The project will start early 2010. This project will continue the open lines of predecessor project KNOW by including the outcomes of KYOTO.

For evaluation, we proposed to organize a task in the SemEval-2010 competition to evaluate the detection of word meanings in the specific domain of the environment. Separate evaluations have been carried out to benchmark the detection of named entities (dates, periods, locations and regions) and first evaluations will be carried out for the extraction of terms and concepts and the detection of facts in the estuary databases. A first evaluation of semantic search will be done as well before the end of the second year. The evaluation platform is ready and we are waiting for the completion of the estuary databases to carry out the evaluation. This will complete the KYOTO knowledge cycle in the project.

The website of the project already has extensive information and demonstrations. It has currently over 20,500 visitors from all over the world. The project has been presented at various conferences and workshops and we created posters and brochure for easy dissemination. All publications, technical papers, project deliverables and presentations can be downloaded from the public website. The website further gives access to demos, videos and tutorials for all major modules and data formats.

7. Future Work or Exploitation Prospects

We will soon release a first integrated system as an open source package that can be downloaded and used by software developers outside the project. The third and final year of the project will further focus on improving and fine-tuning the system through several evaluation and development rounds. This will result in new releases of the system during the year.

The industrial partners IRION and SYNTHEMA will explore the exploitation of KYOTO in new products and services. The current semantic search system is a first demonstration of such an exploitation. Other systems that we foresee are:

1. a FactAlert application that will alert users for changes in factual information that is published in textual form on a specified list of online sources
2. a Dialogue system through which non-experts users can interactively explore knowledge and facts on the environment.

The Wikyplanet system will be used to embed the KYOTO system in an easy to use community platform. The users in the environment domain can see the direct benefit of KYOTO in relation to the knowledge sources that they specify for specific cases. This can be in the form of semantic search options that are directly integrated in Wikyplanet, e.g. browsing or searching facts related to the uploaded sources.

At the end of the third year of the project, we will organize a final workshop to present the final system and the demonstration platform in the environment domain. The workshop will target a wider audience and explore future exploitations and developments of KYOTO.

11. Further Information

Events in connection with KYOTO:

Workshops:

- 1st public KYOTO Workshop, Amsterdam, the Netherlands, February 2-3, 2009 (www.kyoto-project.eu)
- SemEval2010 task on domain specific word-sense-disambiguation
- Global Wordnet Conference 2010, Mumbai India (www.globalwordnet-iitb2010.in)

Demos:

The architecture of KYOTO is shown in the flash animation below. In the schema, the different modules of the KYOTO system are given as rectangular boxes and the data structures by blue repositories. The animation shows two cycles of processes in KYOTO which are explained in more detail below.
The cycles start with documents and websites that are provided by the users in the project: ECNC and WWF.We have collected a first set of documents websites in 4 languages that focus on a series of environmental themes. If you click on the logos of the users, you will get a baseline retrieval system for these documents.

The KYOTO Modules

The KYOTO system has the following modules:

1. **Syntactic processors**: they produce a syntactic and morphological analysis of the text
2. **Semantic processors**: they determine what the meaning is of the words in the text
3. **Tybots**: they learn the terms that are used in the documents and organize these as a hierarchy. If you click on the term extractor module, you can access a demo that gives access to term databases that have been extracted from the environment documents.
4. **Term editor** (part of the Wikyoto platform): users can edit the terms, give definitions and agree on what they mean. These users are called concept users since they are domain experts that maintain the terminology. You can click on the module to go to a demo of the term editor and try it out yourself.
5. **Kybots**: little programs that use the knowledge built up for terms to extract facts from any set of documents. If you click on this module, you will access a demo where you can design or submit a Kybot to extract facts from a sample database
6. **NL Query**: search module with which any end user (people from the domain, government, companies, students, children, etc.) can access the database of facts that is produced. If you click on this module, you can access a demo on semantic search on the English data.

The KYOTO Repositories

In the architecture there are also 4 databases:

1. **Document base**: a database that holds all the documents after being processed by the syntactic and semantic processors. The text is represented in a special XML format

called the KYOTO Annotation Format. If you click on the database, you can see examples of this format in different languages and get access to the DTD.

2. **Term database**: this database holds the output of the term extraction. The terms can be exported into XML in a special format that is called KYOTO-TMF. If you click on the database, you get more details on the term structure in TMF.
3. **Multilingual Knowledge Base**: this database holds the wordnets in all the languages and ontologies that are already given. It holds also any domain wordnet and ontology that is built by editing the term database. If you click on the database, you can view the databases, which are represented in a special XML format for wordnets (Wordnet-LMF) and for ontologies (OWL).
4. **Fact database**: this is the database in which all the extracted facts are stored. This database still needs to be designed in the project. Further details will follow. Note that the database can also hold changing realities. It can extract a fact at some point in time and another fact related to the same things and same place at another point in time.

The KYOTO Cycles

Documents and websites are then processed in two cycles, which is shown in the animation:

1. First cycle in which concept users upload specialized documents and sources to acquire a good term database and to enable them to build a good domain wordnet and ontology:
   1. sources are processed syntactically and semantically and the output is stored in the document base as KAF-XML
   2. the Tybots extract the terms and put the terms in the term database
   3. the domain specialists review and modify the terms, define their meaning and agree with the meanings of terms in other languages through the ontology
2. Second cycle in which the same documents or any other set of documents are sent to KYOTO to extract any facts:
   1. sources are processed syntactically and semantically and the output is stored in the document base as KAF-XML (same as in the first cycle)
   2. the Kybots extract the facts that the end-users are interested in and stores the facts in the fact database
   3. End users get alerts on new facts or can search in the database to get comprehensive and precise informations that can be organized in many different ways, e.g. per region or along time lines, to reveal trends and changes.

Links to the demos:

- Baseline Search
- Multilingual Knowledge Repository
- Terminology extraction: TYBOT
- Wikyoto Term Editor
- Wikyoto Kybot Editor (Fact extractor)
- Fact Retrieval System

PR:

- PR Brochure on KYOTO
- KYOTO participates in the Global WordNet Grid (http://www.globalwordnet.org/gwa/gwa_grid.htm) of the Global WordNet Association (www.globalwordnet.org).

- Opening of a technology forum and an environment forum to open up discussion to a wider public at www.kyoto-project.eu.
- A0-poster on KYOTO, presented at LREC, Marrakech, Morocco, May 28-30, 2008.
- KYOTO used as a promotion/demonstration of present projects combining language and ICT on the "European Day of Languages" for the Representation of the EU in Rome. "EU corner" at the Night of Researchers' exhibition, September 25, 2008
- Factsheet on Tybots and Kybots in Wikyoto

Presentations/Invited talks:

- Piek Vossen: invited keynote speaker on "KYOTO" at eLexicography in the 21st century: new challenges, new applications, Louvain-la-Neuve, Belgium, October 22-24, 2009
- Prof. Dr. PTJM Vossen: Invited Member on "KYOTO"" of Panel in FLaReNet/SILT Workshop on Lexicon-Ontology Relationship in connection with GL 2009 - 5th International Conference on Generative Approaches to the Lexicon, Pisa, Italy, September 17-19, 2009.
- Piek Vossen: Keynote speaker on "KYOTO" at the 25th Annual Conference of the Spanish Society for Natural Language Processing 2009 (SEPLN´09), San Sebastian, Spain, September 8-10, 2009
- Nicoletta Calzolari: invited speaker at the PAROLE Consortium Workshop “New horizons for Linguistic Resources in a Global Context”, presented KYOTO in her talk “From PAROLE to FLaReNet and beyond”, Barcelona 7-8 July 2009.
- Monica Monachini: invited speaker at the PAROLE Consortium Workshop “New horizons for Linguistic Resources in a Global Context”, Barcelona 7-8 July 2009, gave a talk on Standards for Lexixal Resources and presented the KYOTO WordNet-LMF and the KYOTO WordNet Grid as a model for a future PAROLE Grid.
- Piek Vossen: Invited speaker on KYOTO at EU-Japan ICT Cooperation Forum on ICT, Brussels, Belgium, July 2, 2009.
- Nicoletta Calzolari presented KYOTO during the ISO TC47 SC4 Plenary Meeting held in Boulder (Colorado), June 2009.
- Nicoletta Calzolari presented KYOTO during the Institutional Evaluation of CNR-ILC research activities by an International Panel of experts of the field (10 June 2009).
- Piek Vossen: Invited speaker on "KYOTO" at the EU Information and Networking Event for the 5th Call (ICT FP7 SO 4.3) on Intelligent Information Management, Luxemburg, May 12, 2009
- Nicoletta Calzolari: PhD courses at the Pisa University, presentation of KYOTO, Pisa, May 4, 2009.
- Nicoletta Calzolari: Endinburg University, invited talk, Edinburg, April 17, 2009. She presented KYOTO in her speech “Language Resources: from local initiatives to priorities and challenges in the International scenario”.
- Nicoletta Calzolari: CLARIN meeting, presentation of the KYOTO architecture as an exemplary instantiation of interoperability of language resources and tools, Athens, April 4-5, 2009.
- Nicoletta Calzolari: CLARIN meeting, presentation of the KYOTO Annotation Format used for NLP analysis, as an instantiation of an ISO-conformant format for Language Resources to be used as input for the CLARIN infrastructure, Barcelona, April 4-5, 2009.
- Piek Vossen: invited keynote speaker on "KYOTO" at the Asian Language Resource Summit (ALRS), Phuket, Thailand, March 20-21, 2009.

- Piek Vossen: invited guest Lecture "From WordNet to Global WordNet for language technology", Course Language Engineering Applications, Center for Computational Linguistics, University of Leuven, Belgium, February 26, 2009.
- Claudia Soria: Presentation on KYOTO at IWIC 2009, Stanford, United States of America, February 20-21, 2009.
- German Rigau: Brief summary on the KYOTO-project at III JORNADAS PLN-TIMM, Modelos y técnicas para el acceso a la información multilingüe y multimodal en la web, Colmenarejo, Spain. February 4-5, 2009.
- Poster on KYOTO at CAMeRA@VU, VU University, January 30, 2009, Amsterdam
- Wauter Bosma: Presentation on KYOTO: November 10, 2008, University of Twente, the Netherlands
- Piek Vossen: invited keynote speaker on KYOTO at Lustrum of NL TERM, October 25, 2008, Amsterdam, the Netherlands
- Piek Vossen: invited keynote speake on KYOTO at DART 2008: 2nd Workshop on Distributed Agent-based Retrieval Tools, 10 September 2008, Cagliari, Italy
- Brochure on KYOTO
- Chu-Ren Huang: Invited speaker on the KYOTO Project: “Introducing a multilingual view towards global infrastructure in semantic computing and knowledge engineering”. "Semantic Computing" Panel. 2008 AI Forum, Taipei, Taiwan, May 17-18, 2008
- Piek Vossen: invited guest Lecture on Corpusgebaseerd tekstonderzoek for the ICT Onderwijscentrum of the Vrije University Amsterdam, December 13, 2007.

Publications:

- Vossen P., E. Agirre, F. Bond, W. Bosma, C. Fellbaum, A. Hicks, S. Hsieh, H. Isahara, Ch. Huang, K. Kanzaki, A. Marchetti, G. Rigau, F. Ronzano, R. Segers, M. Tesconi (fc.): "KYOTO: a Wiki for Establishing Semantic Interoperability for Knowledge Sharing across Languages and Cultures", in: Handbook of Research on Culturally-Aware Information Technology: Perspectives and Models. Ed. Dr. E. Blanchard (Mc Gill University (Canada)) and Dr. D. Allard (Dalhousie University), IGI Global USA.
- Calzolari, N., Monachini, M., Quochi, V., Soria, C., Toral, A. (fc) Lexicons, Terminologies, Ontologies: "Reflections from Experiences in Resource Construction". To be published in N. Dershowitz and E. Nissan (eds.): Volume in honor of Yaakov Choueka, LNAI Festschrift Volume of the Lecture Notes in Computer Science series, Springer, Berlin.
- Marchetti, A., S. Minutoli, F. Ronzano, M. Tesconi “WIKYOTO Knowledge Editor: The collaborative web environment to manage KYOTO Multilingual Knowledge Base” (to appear) in the Proceedings of the 6th International Conference on Knowledge Management, Hong Kong, China, December 3-4, 2009.
- Agirre, E., A. Casillas, A. Díaz de Ilarraza, A. Estarrona, K. Fernández, K. Gojenola, E. Laparra, G. Rigau, A. Soroa: “Kyoto Project”, in: proceedings of the 25th edition of the Annual Conference of the Spanish Society for Natural Language Processing 2009 (SEPLN´09). Donostia, Spain. 2009.
- **Best Student Paper Award:** Herold, A., and A. Hicks: “Evaluating Ontologies with Rudify”, in: Proceedings of Knowledge International Conference on Knowledge Engineering and Ontology Development, Madeira, Portugal, October 6-8, 2009.
- Wauter Bosma, Piek Vossen, Aitor Soroa, German Rigau, Maurizio Tesconi, Andrea Marchetti, Monica Monachini and Carlo Aliprandi: “KAF: a generic semantic annotation

format", in: Proceedings of the 5th International Conference on Generative Approaches to the Lexicon GL 2009, Pisa, Italy, September 17-19, 2009.

- Toral A., M. Monachini, A. Soroa, G. Rigau: "Studying the role of Qualia Relations for Word Sense Disambiguation", in: Proceedings of GL 09, Pisa, Italy, September 17-19, 2009.
- Toral A., O. Ferrández, E. Agirre, R. Muñoz: "A study on Linking and disambiguating Wikipedia categories to Wordnet using text similarity", in: Proceedings of RANLP 2009, Borovets, Bulgaria, September 14-16, 2009.
- Egoitz Laparra and German Rigau. Integrating WordNet and FrameNet using a knowledge-based Word Sense Disambiguation algorithm. In: Proceedings of RANLP 2009, Borovets, Bulgaria, September 14-16, 2009.
- W. Bosma: "Contextual salience in query-based summarization:, in Proceedings of RANLP 2009, Borovets, Bulgaria, September 14-16, 2009.
- Eric Yeh, Daniel Ramage, Chris Manning, Eneko Agirre and Aitor Soroa: "WikiWalk: Random walks on Wikipedia for Semantic Relatedness", in: Proceedings of the ACL workshop "TextGraphs-4: Graph-based Methods for Natural Language Processing", Singapore, August 7, 2009
- Eneko Agirre, Oier Lopez de Lacalle and Aitor Soroa: "Knowledge-based WSD and specific domains: performing over supervised WSD", in: Proceedings of IJCAI. Pasadena, USA, July 11-17, 2009
- Agirre E., E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, A. Soroa: ""A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches", in: Proceedings of NAACL-HLT 09. Boulder, Colorado, May 31- June 5, 2009.
- Agirre E., A. Soroa: "Personalizing Page Rank for Word Sense Disambiguation", in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009), Athens, Greece, March 30-April 3, 2009.
- Agirre E., Lopez de Lacalle: "Supervised Domain Adaption for Supervised WSD Systems", in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece, March 30-April 3, 2009.
- Izquierdo R., A. Suárez, G. Rigau: "An Empirical Study on Class-based Word Sense Disambiguation", in: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009). Athens, Greece, March 30-April 3, 2009.
- Soria C., M. Monachini, P. Vossen "Wordnet-LMf: fleshing out a standardized format for wordnet interoperability", in: Proceedings of IWIC2009, Stanford, USA, February 20-21, 2009.
- Cuadros, M., G. Rigau. "KnowNet: Building a Large Net of Knowledge from the Web", in: Proceedings of COLING2008, Manchester, 18-22 August, 2008.
- Vossen P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon (2008). "KYOTO: A system for Mining, Structuring and Distributing Knowledge Across Languages and Cultures", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.
- Agirre E. and Soroa A., "Using the Multilingual Central Repository for Graph-Based Word Sense Disambiguation", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.
- Álvez J., Atserias J., Carrera J., Climent S., Laparra E., Oliver A. and Rigau G. "Complete and Consistent Annotation of using the Top Concept Ontology", in: Proceedings of LREC 2008, Marrakech, Morocco, May 28-30, 2008.

- Pociello E., Gurrutxaga A., Agirre E., Aldezabal I. and Rigau G. "WNTERM: Enriching the MCR with a terminological dictionary", in: Proceedings of LREC2008, Marrakech, Morocco, May 28-30, 2008.
- Vossen, P., E. Agirre, N. Calzolari, C. Fellbaum, S. Hsieh, C. Huang, H. Isahara, K. Kanzaki, A. Marchetti, M. Monachini, F. Neri, R. Raffaelli, G. Rigau, M. Tescon, J. van Gent (2008): " KYOTO: A System for Mining, Structuring, and Distributing Knowledge Across Languages and Cultures ", in: Proceedings of the Fourth International Global Word Net Conference - GWC 2008, Szeged, Hungary, January 22-25, 2008

Deliverables :

D 11.1	Project Public Website	Final	P	March 2008
D 1.1	Data types and sources	1.0	P	May 2008
D 7.1	XML Schema for Wordnet and Ontology	2.0	P	May 2008
M03	M03 Quarterly Management Report	Final	R	June 2008
D 1.2	Knowledge types and questions	1.0	P	August 2008
D 2.1	Database models and data formats	3.2	P	August 2008
D 2.2	Overall system design	2.0	P	August 2008
M06	M06 Quarterly Management Report	Final	R	September 2008
D 9.1	User-scenarios		P	October 2008
D 7.2a	Knowledge Base Server API version 1	1.0	P	November 2008
D 3.1	Capture module	1.0	P	November 2008
D 6.1	Accumulated knowledge	1.0	P	November 2008
AR	Annual public report 2008	Final	P	November 2008
M09	M09 Quarterly Management Report	Final	R	December 2008
D 1.3	User Requirements (draft)	2.0	P	March 2009
D 11.2	1st Kyoto Workshop	Final	P	March 2009
D 7.4a	Wiki Environment for WordNet Editing (draft)	1.0	P	March 2009
D 4.1	Indexing Module (draft)	1.0	P	March 2009
D 11.5	Design of Showcase DVD (draft)	1.0	P	March 2009
D 6.2	Central Ontology (draft)	3.0	P	March 2009

D 11.4	Early Dissemination Plan & Strategy (draft)	4.0	P	March 2009
M12	M12 Quarterly Management Report	Final	R	March 2009
D 6.3	Wordnets mapped to Central Ontology (draft)	2.0	R	March 2009
D 5.1	Concept Miners version 1 (draft)	2.0	P	March 2009
D 7.3a	Multilingual Wordnet Services (draft)	2.0	P	March 2009
D 7.5a	Wiki Environment for Ontology Editing (draft)	1.0	P	March 2009
M15	M15 Quarterly Management Report	1.0	R	June 2009
D 5.2a	Fact Miners Version 1	1.0	P	August 2009
D 9.2	Search Engine Client	1.0	P	August 2009
M18	M18 Quarterly Management Report	1.0	R	August 2009
D 6.4	Automatic Deduction and Inferencing Technques	1.0	P	August 2009
D 8.1	Domain Extension of Central Ontology - version 1	1.0	P	August 2009
D 8.2	Domain Extension of Wordnets - version 1	1.0	P	August 2009
D 9.3	Benchmark Test of 1st Indexes	1.0	P	September 2009
	Annual Report 2009	1.0	P	November 2009
D 7.2b	Knowledge Base Server API - revised	1.0	P	November 2009
D 9.4	End-user Test of 1st Indexes	1.0	P	November 2009
D 7.3b	Multilingual Wordnet Services - revised	1.0	P	December 2009
M21	M21 Quarterly Management Report	1.0	R	December 2009

Working Papers:

- TR002/WP02 KYOTO LMF WordNet Representation Format, v.04
- TR006/WP02 Fact Annotation Format for KYOTO
- TR002/WP02 KYOTO LMF WordNet Representation Format, v.03
- TR004/WP02 KYOTO: The representation of terms, v.01
- TR003/WP02 Formalizing Knowledge by Ontologies: OWL and KIF, v.02
- TR001/WP01 User Scenarios Wikyoto, v.01
- TR005/WP05 Storyboard: "to mine by example" for building Kybots, v.01