KYOTO

Project acronym: ***KYOTO***
Project full title: ***Knowledge Yielding Ontologies for Transition-based Organization***
Grant agreement no.: ***211423***

Project summary

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 and other environmental issues related to rapid growth and economic developments. The globalization of problems and their solutions requires that information and communication be supported across a wide range of languages and cultures. Specifically, we need a system that is able to collect and represent in a uniform way distributed information structured and expressed differently across languages. Such a system should furthermore allow both experts and laymen to access this information in their own language, without recourse to cultural background knowledge. Natural Language is the most ubiquitous and flexible interface between users -especially non-expert users- and information systems. 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 goal of KYOTO is to develop a content enabling system that provides deep semantic search and information access to large quantities of distributed multimedia data for both experts and the general public, covering a broad range of data from wide-spread sources in a number of culturally diverse languages. In this project we will target the languages: English, Dutch, Italian, Spanish, Basque, Chinese and Japanese. This powerful system crucially rests on an ontology linked to wordnets—lexical semantic databases--in a variety of languages. Concept extraction and data mining are applied through a chain of semantic processors that re-use the knowledge for different languages and for particular domains. The shared ontology guarantees a uniform interpretation for diverse types of information from different sources and languages. The system can be maintained by field specialists using a Wiki platform. KYOTO is a generic system offering knowledge transition for any domain of knowledge and information, across different target groups in society and across linguistic, cultural and geographic borders. KYOTO will be applied to the environmental domain and span global information across European and non-European languages.

Concept and project objectives

The goal of KYOTO is to develop a content enabling system that provides semantic search and information access to large quantities of distributed multimedia data for both experts and the general public, and to apply this system to environmental information on a global scale. Information access is provided through a cross-lingual user-friendly interface that allows for high-precision search and information dialogues over a variety of data from wide-spread sources in a range of different languages: English, Dutch, Italian, Spanish, Basque, Chinese and Japanese This is made possible through a customizable ontology that is linked to various wordnets—lexical databases with rich semantic information--and a set of knowledge yielding text miners (so-called Kybots) for a variety of languages. Concept extraction and text mining is applied through a chain of linguistic and semantic processors that share a common ground and knowledge base. The shared ontology guarantees a uniform interpretation layer for the diverse information from different sources and languages. The system can be maintained and kept up to date by specialists in the field using an open Wiki platform for ontology maintenance and wordnet extension. KYOTO is a generic system that offers knowledge transition across different target groups in society and across linguistic, cultural and geographic borders. In this project, the KYOTO system will be applied to the environmental domain and span information on a global scale across European and non-European languages. KYOTO addresses the need for global and uniform transition of knowledge across different types of organizations, which is particularly critical in the environmental domain.

Figure-1 below gives a schematic overview of the complete system. At the top of the diagram, a collection of source data in different media and languages is given. Global environmental organizations will supply access to the information sources and specify the types of knowledge that should be disclosed. The aim is to build databases that allow one to find:

- web pages and websites with specific information
- human experts that can help in specific situations and address specific needs
- documents and specific paragraphs with answers
- pictures, tables and schemes
- specific facts

In collaboration with the end-users, additional types of information can be defined. At this point, we have commitments from numerous organisations to provide access to their data and that have an interest in the solution, among them Greenpeace International, WWF, Mooi Informatie Beheer. WWF is a subcontractor for the project; other interested organizations will be included in a user group.

The sources for the information can be registered to a capture module that will collect the information and produce a general XML representation for the data from each source. The information in the XML representation is processed by a chain of linguistic and conceptual processors. Through wordnets in each of the languages, the textual information will then be matched to a shared universal ontology. This ontology guarantees a common level of semantic anchoring across languages and information sources.

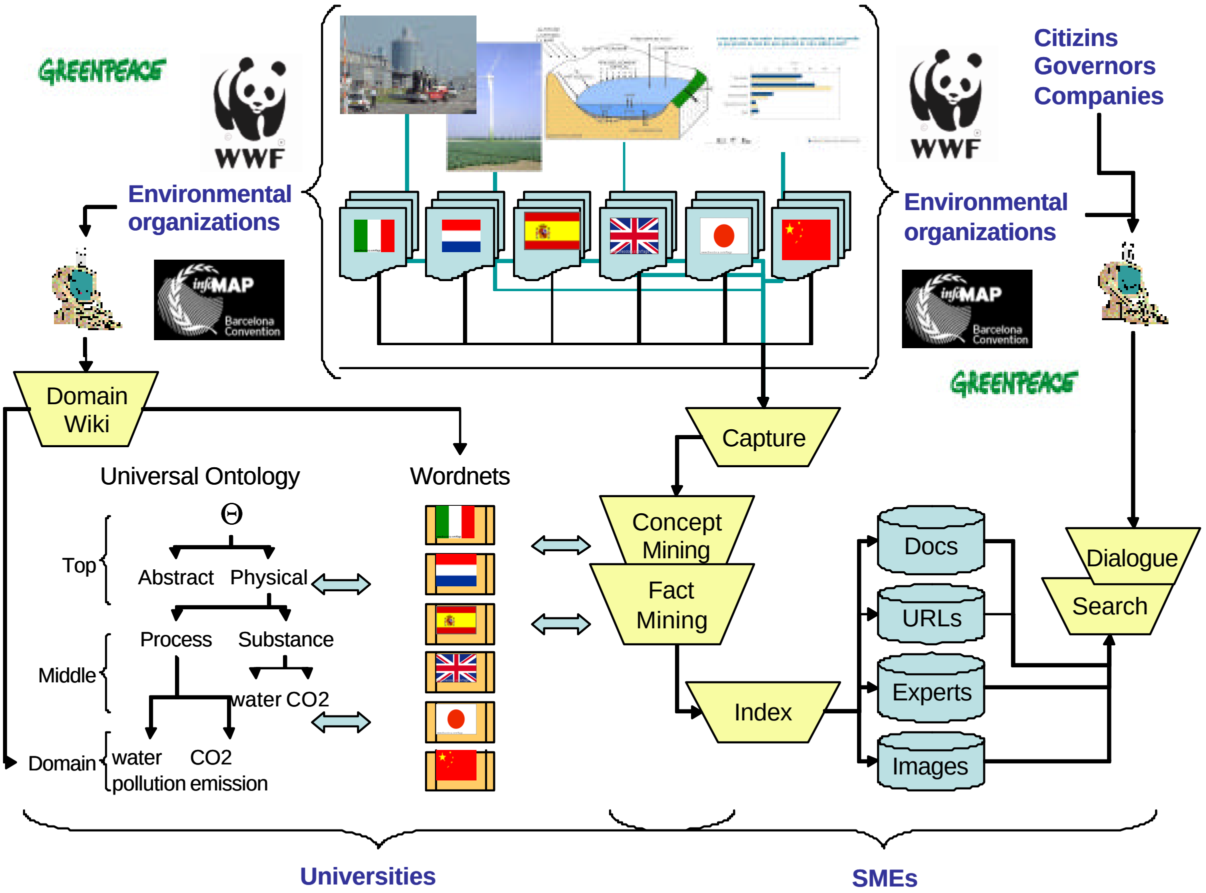

Figure 1: System architecture

The ontology consists of three layers. The top layer will be based on existing top level ontologies, among them SUMO (Niles and Pease 2001 and 2003) and DOLCE (Masolo et al 2003). An ontology is a definition of knowledge in a formal logical format. The middle-layer will be derived from the existing wordnets. Wordnets are databases with words from languages mapped to concepts. See the next section for a short explanation of wordnets and ontologies.

The middle level needs to be developed to connect the domain terms and concepts to the top-level.[1] The domain terms are extracted semi-automatically from the source documents but are also manually created through a Domain Wiki. The Domain Wiki lets experts in the field modify and extend the domain level of the ontology and extend the corresponding wordnets in each language. It enables community-based resource building which will lead to better understanding and consensus in the field and at the same time result in the formalization of this knowledge so that it can be used by a system. Extensions to wordnets and the ontology are propagated to other wordnets and language resource builders through a sharing protocol: XFlow (Marchetti et al. 2006), and LexFlow (Tesconi et al. 2006, 2007, Soria et al. 2006a,b).

Once the ontological anchoring is established, it will be possible to build text mining software that can detect semantic relations and facts in text. These data miners, so-called Kybots (Knowledge Yielding roBots), can be defined using constraints between relations at a generic

[1] Within the scope of the project this work is necessarily limited

ontological level. These logical expressions need to be implemented in each language by mapping the conceptual constraint on linguistic patterns. A collection of Kybots created this way can be used to extract the relevant knowledge from textual sources associated with various media and genres and different languages and cultures, and represent the result in a uniform and standardized XML format, compatible with WWW specifications for knowledge representation such as RDF and OWL.

The extracted knowledge and information will be indexed by a corporate search system that can handle fast semantic search across languages. The search system uses so-called contextual conceptual indexes, which means that occurrences of concepts in text are interpreted by their co-occurrence with other concepts within a linguistic context, e.g. a noun phrase, sentence or some other pattern. The co-occurrence relation of concepts can be specified in various ways, possibly being based on semantic relations that are defined in the logical expressions. Likewise the search system can give different results searching for "polluting substance" than for "polluted substance", because these involve different concepts and semantic relations. By mapping a query to concepts and relations, very precise matches can be generated, without the loss of scalability and robustness found in regular search engines that rely on string matching and context windows.

The indexing, search and interfacing processes are based on existing commercial systems. Innovation is based on the conceptual richness and detail of information that is disclosed in the KYOTO environment. Reasoning over facts and ontological structures will make it possible to handle diverse and more complex types of questions. Understanding across languages and cultures is established through the ontological anchoring of language via wordnets and text miners.

The KYOTO system represents knowledge sharing in 4 different aspects:

1. Generic knowledge is re-used and shared in various domains
2. Generic Kybots (knowledge yielding miners) are re-used and shared in various domains
3. Ontologies (both generic and domain-specific) are shared across languages
4. Kybots (both generic and domain-specific) are re-used and shared across languages

In Figure 2 a schematic representation is given of the sharing and interoperability of the knowledge in KYOTO. Sharing of generic ontological knowledge to the domain will mainly take place through subclass relations. We will collect all the relevant terms in each language for the domain and add them to the general ontology. Possibly, these concepts can be imported from a specific wordnet and ontologized. It will be important to specify exactly the ontological status of the terms. Only disjunct types need to be added (Fellbaum and Vossen 2007). For example, ***CO2*** is type of substance, but ***green-house gas*** does not represent a different type of gas or substance but refers to substances that play a specific role in specific circumstances. In so far as new definitions and axioms need to be specified, these can be added for the specific subtypes in the domain. However, this is only necessary if the related information also needs to be mined from the text and is not already covered by the generic miners. Next, the generic and domain knowledge is shared among all participating languages through the mapping of the different wordnets to the ontology.

Sharing of Kybots will be more subtle. For example, *concentrations of substances*, *causal relations between processes* or *conditional states for processes* can be stated as general conceptual patterns using a simple logical expression. Within a specific domain, any of these relations and conditions could be detected within the textual data by just using these general patterns. For example, people usually do not use special words in a language to refer to the causal relation itself but they use general words such as "cause" or "factor". Since any causal relation may hold among processes and or states, they can also hold in the environmental domain. Certain valid conditions can be specified in addition to the general ones, as they are relevant for the users. For example, CO2 emissions can be derived from a certain process involving certain amounts of the substance CO2 but critical levels can be defined in the text miner as a conceptual constraint. Furthermore, we may want to limit the ambiguity of interpretation that arises at the generic levels to only one interpretation at the domain level. To what extent generic patterns can be used or need to be tuned will be investigated in the project.

Each language group can then build a Kybot for their language, capturing a particular relation. They can either re-use a given logical expression that underlies the Kybot of another language, or they can formulate a new pattern in their language and derive a generic universal pattern from it. For example in Figure 2, a generic linguistic text miner is formulated for Dutch, based on Dutch words and expressions. This Kybot is projected to the ontology via the Dutch wordnet, becoming a generic ontological expression which relates two ontological classes: a Substance to a Process. This expression may be extended to a domain, where it is applied to CO2 and CO2 emissions. Next, the Spanish group can load the domain specific expression and transform it into a Spanish Kybot that can be applied to a domain text in Spanish. To turn an ontological expression into a Kybot, language expressions rules and functions need to be provided. This process can be applied to all the participating languages, where the basic knowledge is shared.

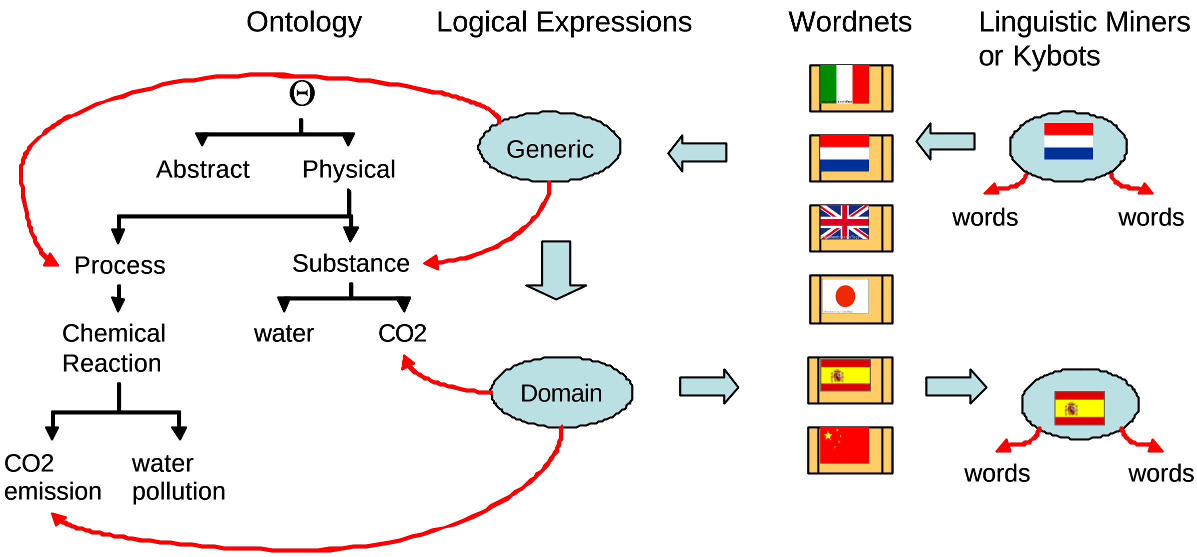

Figure 2: Levels of sharing and interoperability

KYOTO will thus generate Kybots (processors) in each language that go back to a shared ontology and shared logical expressions. Likewise, KYOTO can be seen as a sophisticated platform for anchoring and grounding meaning within a social community, where meaning is expressed and conceived differently across many languages and cultures. It also immediately makes this shared knowledge operational so that factual knowledge can be mined from unstructured text in domains. KYOTO supports interoperability and sharing across these communities since much knowledge can be re-used in any other domain.

To summarize, the concrete objectives of KYOTO are:

1. An open knowledge sharing and anchoring system.
2. Ontologies: We will define all the high-level and mid-level concepts that are needed to accommodate the information in the environmental domain. Knowledge is implemented at the most generic level to maximize the re-usability but still precisely enough to yield useful constraints in detecting relations. Within the domain, we will extend the ontologies to cover all the necessary concepts and relations that apply and can be shared. The database that holds the ontologies and the XML data will be made available for free for the whole community.
3. Wordnets: existing wordnets will be extended and harmonized given the ontology developed in the project. The database that holds the wordnets and the XML data (content) itself will be made freely available for the whole community.
4. Acquisition tools: we will develop software in all 7 languages to automatically extract synsets and synset-relations from text within a domain.
5. Linguistic processors: we will use linguistic processors in each language to carry out basic analysis: tokenization, segmentation, tagging, parsing and word-sense disambiguation. Where possible we will use existing technology and resources.
6. Kybots will be developed to cover the questions and answers that are listed by the users and to cover generic concepts and relations that occur in any domains, such as named-entities, locations, time-points, etc. Kybots are primarily defined at a generic level to maximize re-usability and inter-operability. The specification of the kybots and the software will be made available for free for the whole community. We will develop the kybots that are necessary for the selected domain and the types of question and knowledge that the users have defined. However, the system can easily be extended and ported to other domains and we will make plans for this in the exploitation work package.
7. Concept and knowledge representation formats will be defined to store the data and index them, compatible with given standards (ISO and semantic web standards)
8. The Wiki environment will have some general characteristics typical of a generic wiki engine such as:

    - an easy interface tailored to domain experts who don't know the underlying complex data model (ontology plus multi grid wordnet);
    - a simplified wiki syntax that is much easier to use for non technical users than e.g. HTML;
    - a web based interface;
    - a rollback mechanism: each change to the content is versioned;
    - search functions: synset;
    - automatic downloading of information from web resources e.g. Wikipedia;
    - a support for collaborative editing and consensus achievement such as discussion forums, and list of last updates.
    - a role based user management;

    In addition, the wiki engine has to manage the underlying complex data model in order to keep it consistent. For instance when a new domain term such as "water pollution" is inserted into a language specific wordnet by a domain expert, a new entry will be automatically inserted in the ontology extension and in every wordnet. The Wiki environment will list all dummy entries to be filled in. English may be used as the common ground language in order to support the extension process and the propagation of changes between the different wordnets and the ontology.

9. Portal: the cross-lingual portal will be used to shows the effects of deep semantic processing to the community and to match queries across languages and cultures. The system is based on proprietary software but the knowledge levels are open and can be used by any other system.

All the results of KYOTO will be public and freely available.

Table 1: Consortium

Partner no.	**Partner's name**	**Partner's short name**	**Country**
1 (Coordinator)	Faculteit der Letteren, Vrije Universiteit Amsterdam	VUA	NETHERLANDS
2	Consiglio Nazionale delle Ricerche	CNR-ILC-IIT	ITALY
3	Berlin-Brandenburg Academy of Sciences and Humantities	BBAW	GERMANY
4	Euskal Herriko Unibertsitatea	EHU	SPAIN
5	Academia Sinica	AS	TAIWAN
6	National Institute of Information and Communications Technology	NICT	JAPAN
7	Irion Technologies	IRION	NETHERLANDS
8	Synthema	SYNTHEMA	ITALY
9	European Centre for Nature Conservation	ECNC	NETHERLANDS