KYOTO Annual Report 2008

“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. 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.
2. The text in various languages is collected from the sources.
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*.

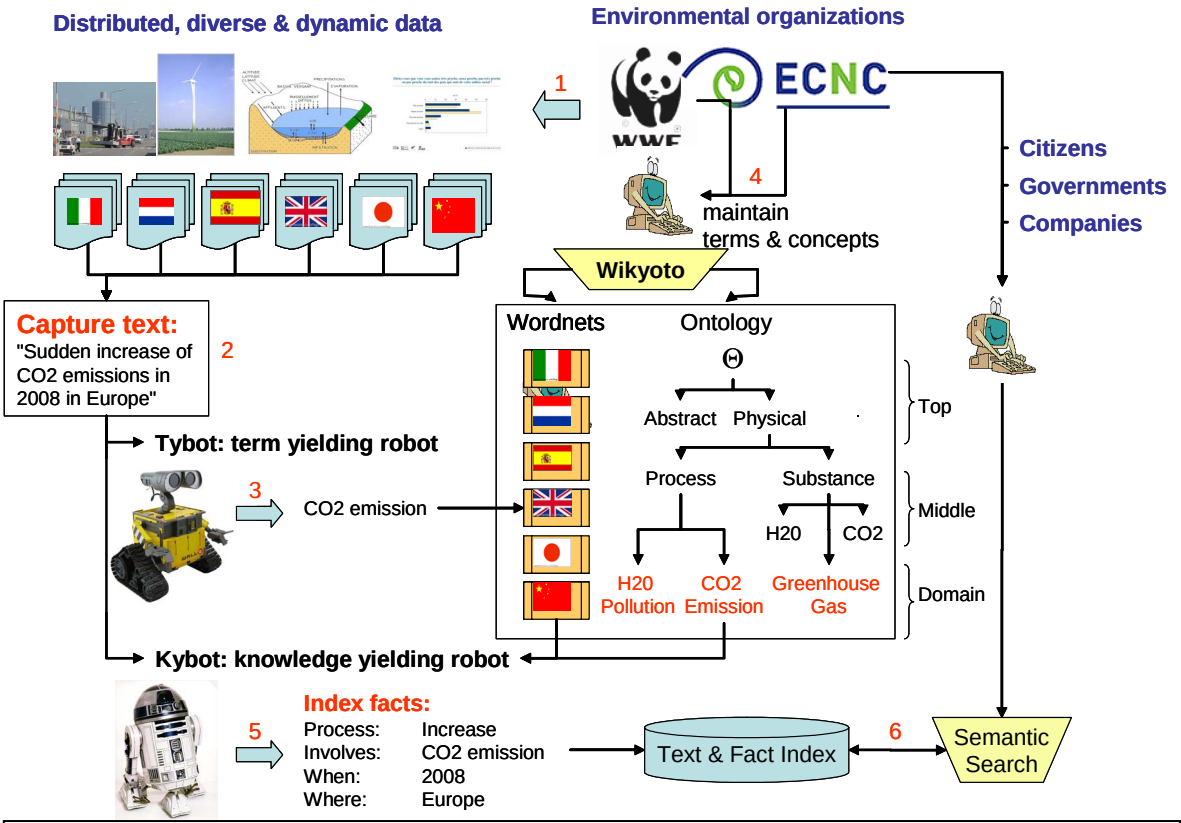

Figure 1: Overall overview of the KYOTO system

2. Summary of Activities

The project is now 9 months old. We carried out a first intense round of defining the user-requirements with a focus on the type of questions and knowledge that the environmental users are interested in. The users defined 7 areas of interest related to ecosystem services: considering the planet and nature as a valuable asset in our society. For these areas, about 350 questions have been formulated that represent the major types of knowledge and facts of interest. Furthermore, they provided over 15,000 sources of documents and websites that may contain this information.

We built an early demonstrator of a baseline system for the source data covering 4 different languages. The system logs all the user actions so that we see what questions and what pieces of text have been used to compile answers in the domain. This represents the basis for evaluating the knowledge mining and semantic search systems built in KYOTO.

We also completed the design phase of the project and are developing the first versions of the modules. The KYOTO system is designed as an open system that can easily be extended and exploited for many languages in the world. For this purpose, we have used (ISO-)standards where possible and partly designed our own specifications. We defined 3 innovative modules:

1. A wiki platform for knowledge experts to model the concepts and terms in their domain, generating formal expressions if this knowledge that can be used by computer software. This system is called the Wikyoto system.
2. Software for automatically extracting the terms and concepts from any document collection. This system is called the Tybot service: Term Yielding robot.
3. Software for automatically detecting facts and semantic relations in text, using the modeled knowledge and terms in the domain. This system is called the Kybot service: Knowledge Yielding robot.

Early 2009, we will deliver the first version of the KYOTO system: KYOTO-1 and carry out the first experiments with the users to evaluate the system. KYOTO-1 will be specialized to extract vast amount of facts on countings of species and populations in areas in the world. These countings are the basic facts for the environment domain. We will also develop KYOTO-2 in 2009, which will focus on changes in populations, causal relations and time-expressions. KYOTO-1 and KYOTO-2 will generate vast amounts of facts for the users that cannot easily be collected by humans, certainly not on a regular basis and across different languages. Various versions of the system will be developed and tested in evaluation frameworks that are set-up.

In the next sections, we will summarize the major areas of work in more detail.

3. User requirements

ECNC and WWF (WNF) have first defined their topics, according to their workfields.

For ECNC the topics are:

- *Nature and Society*
- *Business and Biodiversity*
- *State of Nature and Biodiversity*
- *Ecological Connectivity*

For WWF the topics are:

- *State of Nature and Biodiversity*
- *Ecosystem Services*
- *Nature and Poverty*

Both organizations provided for each topic a list of keywords and questions, generated through a series of facilitated brainstorm sessions, and in case of WWF also by a round of interviews. At the moment the "knowledge types and questions" file contains 630 keywords and 365 questions.

Both organizations also produced a representative set of data types and sources, containing websites and reports, arranged by topic. At the moment there are roughly 500 links in the file, pointing to a range of websites and documents for the Kyoto system to utilize. The data consists of different languages: English, Dutch, Italian, Spanish, Basque, Simplified Mandarin Chinese and Japanese.

The Kyoto system will be developed in 3 different stages. In order to compare (and evaluate) the different stages, a baseline system (or benchmark system) is necessary. This benchmark system is the first demo. The users are working with this demo to derive the user-scenarios.

4. Baseline systems and evaluation framework

A baseline retrieval system has been developed that allows you to search in about 15,000 documents and websites in English, Dutch, Spanish and Italian. The system can be accessed at: http://demo.irion.nl/kyoto. It has been set up in such a way that we can track all the user actions to find the relevant fact for answering complex queries. The users fill in the high-level question that they try to answer and the answer that they have been able to compile. To collect the data, the users can fire queries in a regular search engine. Queries can be formulated in 4 different languages. The results of the retrieval can be viewed and feedback can be given.

Using this system, we monitor how well and efficient users can find the relevant data. The detailed logs are used to derive information on good and bad results, the number of actions needed and the time needed to compile their answers. We also can compare the answers given to the same questions. Future systems of KYOTO will be compared in performance with respect to this baseline system to give us an objective evaluation.

5. System Design

The definition of the overall KYOTO architecture implied the specification of its components and processes, in two different variants, the KYOTO core system (see Figure-2) and the KYOTO+ system. The core system will be free and is released as an open-source system. The KYOTO+ system is extended with background modules to demonstrate a full implementation in which the knowledge can easily be exploited by the end-users.

The basic components are Capture Server, Wikyoto, Tybot and Kybot servers. Since the architecture heavily relies on data formats exchanged among the components, we have investigated available state-of-the-art XML specifications for the various data formats and, in some cases, they have been adapted.

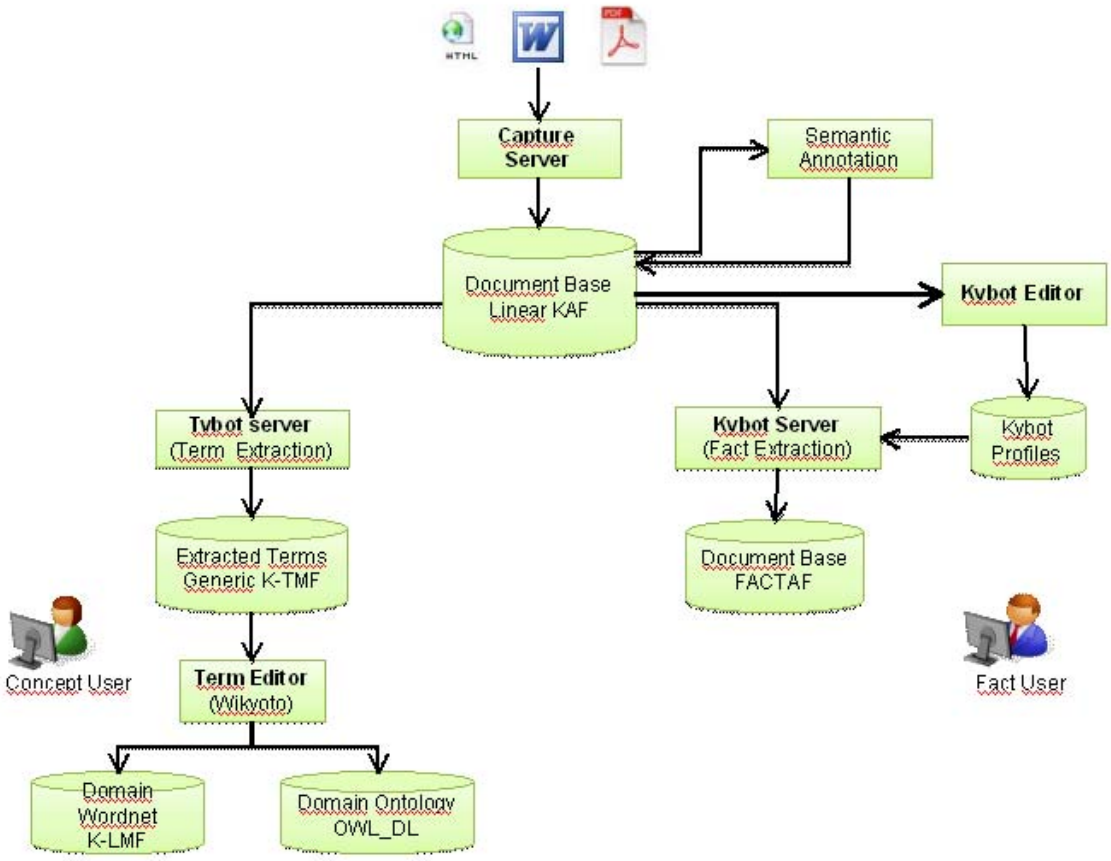

Figure 2: Kyoto Data Flow

We defined a novel multilayered format, Kyoto Annotation Format (KAF) for encoding the output of the Linguistic Annotators. KAF is used to encode Terms, Concepts and Facts extracted from the source documents. We specifically conceived KYOTO-LMF, a dialect of ISO LMF (http://www.lexicalmarkupframework.org/), as a standardized interoperability format for Wordnets. This will pave the way to endowing the different Wordnets with a format that will allow easier integration, with particular reference to the realization of the Global Wordnet Grid.

6. Wikyoto system

We have defined the architecture and the main usage patterns of the Wikyoto system: it represents the global gateway of the KYOTO Knowledge Base. Wikyoto enables social groups of knowledge experts to easily maintain and enrich the linguistic and ontological net of resources as well as to provide inputs to improve the effectiveness of the fact mining robots, called Kybots. The whole Wikyoto system is intended to be browser-accessible via Web: we have thus developed some prototypical Web interface useful to carry out the fundamental tasks supported by Wikyoto to test and better define the user interactions. In particular we have focused our attention on:

- the browsing and enrichment of Wordnets;
- the selection of text patterns relevant to the users to improve Kybots processing effectiveness;
- the editing of relevant terms hierarchies extracted from processed documents in order to enrich the Wordnets and the ontology.

7. Tybots, term yielding robots

One of the goals of Kyoto is to allow communities to build consensus on terminology. Terminology comprises the terms which are used to describe important concepts within the domain as well as how the terms relate to each other. A good view on the domain terminology will be essential for extracting and searching facts.

Both the set of terms and their relations can be extracted automatically to a large extent from a domain corpus. The Tybot (Term Yielding Robot) is responsible for this. For use in Kyoto, a Tybot has been developed which processes text documents to extracts sequences of words which are potentially domain terms. The Tybot also applies hierarchical relations between them. The Tybot operates language independently but it demands that the input documents are processed by a part-of-speech tagger and a chunker. Future work on Tybots includes finding alternative ways to distinguish domain terms from non-terms as well as finding appropriate procedures for evaluating extracted terms and relations.

8. Kybots, knowledge yielding robots

Concept extraction (perfomed by **Tybots**) and text mining (performed by **Kybots**) is applied through a chain of linguistic and semantic processors that share common formats and knowledge bases.

Once the ontological anchoring is established by the **Tybots**, it is be possible to build text mining software that can detect semantic relations and facts occurring among concepts already integrated into the ontologies. These text miners or Kybots (Knowledge Yielding roBots), will be defined by linguistic patterns and semantic constraints expressed at an ontological level. For example, the ontology will give us the conceptual pattern that Populations consist of species that live in a habitat in some region. This information can be realized through e.g. compounding as in: Mediterranean spider population, or as a sentence as in: Large groups of alien spiders that live in dry areas in Mediterranean mountain areas.

In fact, the Kybots will provide a mapping between the conceptual constraints and the linguistic patterns. The output of English, Dutch, Italian, Spanish, Basque, Chinese and Japanese Linguistic Processors will be represented in KAF (Kyoto Annotation Format), a uniform and standardized XML format (see Deliverable D2.1 Database models and data formats).

The facts of interest are defined in so-called Kybot profiles. The profiles can be defined in advance or by individual users. We also designed an initial scenario for the Kybot profile construction we called "Mining by example" (see Working Paper WP05-TR005 for details). In this scenario, the user is provided by an advanced interface allowing for the construction of Kybots from analysed corpus examples, without the need to access the complex conceptual patterns and the linguistic structures.

Collections of Kybots created this way will be applied to extract relevant knowledge from textual sources in different languages and cultures. The Kybots will produce enriched KAF outputs, incorporating (partial) facts.

We will define two early version of the KYOTO system: KYOTO-1 and KYOTO-2. We decided that system KYOTO-1 will focus on a subset of the data:

- species
- populations
- quantities of species
- sizes of populations
- regions

KYOTO-2 can then be an extension to changes related to time and place and causal relations. We planned to generate early output for KYOTO-1 in early 2009. The output will make clear how far we expect to get with Kyoto in terms of coverage and quality of data. This is necessary to further design the Wiki system in WP07.

9. 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.

Furthermore, the project involves two industrial partners: Irion Technologies (Delft, The Netherlands) and Synthema (Pisa, Italy) that develop end-user software systems that heavily rely on language-technology and semantic processing. The industrial partners are capable of developing commercial software products based on the KYOTO system that can be taken to the market relatively soon. One such an application is called the FactAlert system. This application lets end-users define the facts of their interest and the possible sources for finding these facts. The system then scans these sources on a regular basis to see if new facts (changes with respect to the initial fact database) are reported. The user is then informed about these changes as new facts and can directly access the sources for further information.

The consortium also is represented by the Global Wordnet Association (GWA), 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 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). 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. Finally, we submitted a proposal for a Mitsui project (http://www.mitsui.co.jp/en/csr/contribution/activity/index.html) that will exploit the KYOTO system for creating awareness for environmental issues in China. The Mitsui environment initiative is a Japanese fund for Asian projects that address environmental issues. This project, if awarded, will further strengthen the global impact of the project.

For the visibility of the project, we launched a website that gives access to early demos and includes two forums: one for environmental discussions related to the KYOTO system and one for technical issues related to concept and knowledge mining.

The project has been presented at various conferences and workshops and we created a poster and brochure for easy dissemination. All publications, technical papers, project deliverables and presentations can be downloaded from the public website.

10. Future Work or Exploitation Prospects

In the next phase of the project, we foresee the development of the first version of the system that includes the major components. This version can model the complete process flow but the output of the components is still limited in quality, coverage and scope. It will give us however a good basis for improving the individual models through experimentations and user-feedback. Further developments are targeted to include the languages Basque, Chinese and Japanese in the process and to extend the scope and coverage of knowledge that is handled.

We have also planned a first project workshop on February 2nd and 3rd at which the first version of the system will be presented and discussed. The workshop targets both environmental groups and technology developers, raising a discussion on the usability and feasibility of the project, using state of the art technology. The participants are invited to use the early demonstrators and to give direct feedback. We are targeting an audience of 50 experts in both areas.

11. Further Information

Events in connection with KYOTO :

Workshops and demos

- Planned: 1st public Kyoto Workshop, Amsterdam, the Netherlands, February 2-3, 2009 (www.kyoto-project.eu)
- March 1, 2008: 1st Demo launched on Cross-Lingual Search, updated in October 2008 (http://demo.irion.nl/kyoto/)

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).
- Launching 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:

- Prof. Dr. P. Vossen: invited key-note speaker on Kyoto at Lustrum of NL TERM, October 25, 2008, Amsterdam, the Netherlands (Presentation NL Term)
- Prof. Dr. P. Vossen: invited key-note speaker on Kyoto at **DART 2008:** 2nd Workshop on Distributed Agent-based Retrieval Tools, 10 September 2008, Cagliari, Italy (Presentation DART)

Publications:

- 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.
- 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 :

- D1.1 : Data types and sources
- D1.2 : Knowledge types and questions
- D2.1 : Database models and data formats
- D2.2 : Overall system design
- D7.1 : XML Schema for Wordnet and Ontology
- D11.1 : Project Public Website

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