PERIODIC REPORT 3 KYOTO, ICT 211423
version 5

26 April 2012

Editor:
Prof. Dr. Piek Th.J.M. Vossen, VUA, p.vossen@let.vu.nl

Knowledge Yielding Ontologies for Transition-based Organization ICT 211423

PROJECT PERIODIC REPORT
KYOTO

Grant Agreement number: ICT-211423

Project acronym: KYOTO

Project title: Knowledge Yielding Ontologies for Transition-based Organization

Funding Scheme: Collaborative Project -- Small or medium-scale focused research project (STREP)

Date of latest version of Annex I against which the assessment will be made:
21 November 2007

Periodic report: 1st ☐ 2nd ☐ 3rd ■ 4th ☐

Period covered: from March 1, 2010 to February 28, 2011

Prof. Dr. Piek T.J.M. Vossen, VU University Amsterdam, the Netherlands

Tel: +31 (0) 20-5986457

E-mail: p.vossen@let.vu.nl

Project website address: www.kyoto-project.eu

1 Publishable summary

"Knowledge Yielding Ontologies for Transition-based Organization"

website: 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.

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* may 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, their company or their personal footprint.

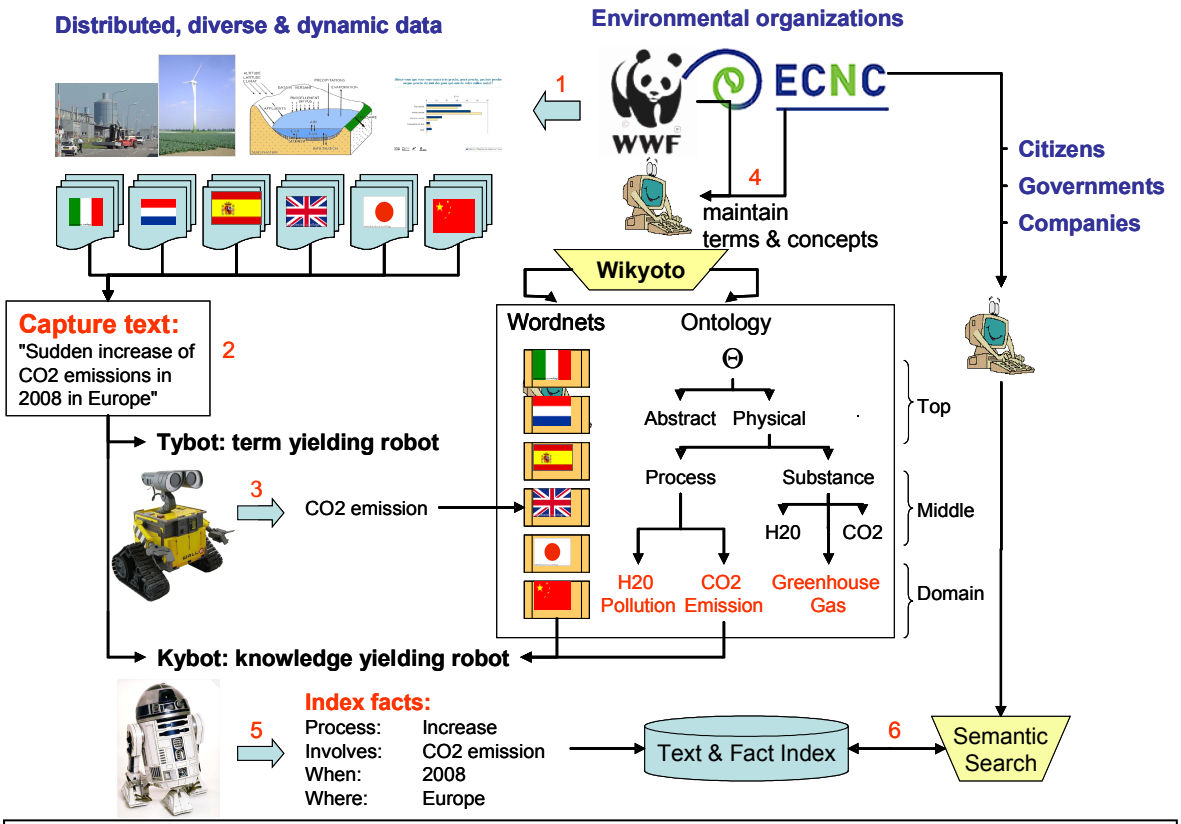

Figure 1: Overall overview of the KYOTO system

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. Text in various languages is collected and captured from these sources, for example the phrase “Sudden increase of CO2 emissions in 2008 in Europe”.

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. From the above sentence, the Tybot will for example learn the term “CO2 emission” as a type of emission that involves the substance CO2 as a by-product of processes. The Tybot will also learn concepts such as *water pollution*, implying that it is the water that is polluted with other substances, or greenhouse gas, which is not a type of gas but gaseous substances with a certain role in global warming. Term extraction is done for multiple languages, which get connected through shared definitions of concepts.

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. It are thus the people in the community that will control the definition of their vocabulary over time, establishing a proper basis for interpreting their language. The Wikyoto system is specifically built so to make it easy for the community people to define the semantic implications of their terms. The domain extensions are nevertheless anchored to the general

vocabulary and ontologies provided. This enables interoperability of the knowledge of the field that is built up by the community.

5. Knowledge yielding robots (so-called **Kybots**), use the terms and knowledge to detect factual data in the text in various languages. A Kybot now knows that certain substances can pollute water and therefore it will be capable to look for such cases in the text. In fact, it can accumulate a comprehensive list of all substances that are mentioned to do so, from massive amounts of text in many different languages. Moreover, the Kybot will determine the time points and regions of such pollutions as well, so that cases of pollution can be depicted on a map or time line. Similarly, Kybots can detect all references to CO2 emissions, related to places and time, and also detect the processes that have led to such emissions.

6. The output of the Kybot is a rich semantic data structure. The Kybots can run on any data set, at any moment of time. Likewise, we can generate these rich structures at different points in time to discover trends or important changes in information. To be able to access this information, the factual data is indexed and special interfaces are developed to search through the knowledge for example for *facts on CO2 emission in Europe from 2000 to 2009*, or *cases of water pollution in your neighborhood*. The semantic architecture allows to organize these data in many new ways (e.g. regional maps, and or timelines) that provide much added value to the end-users, compared to regular searches.

Using the above steps, a sentence such as the following:

Scientists predict that climate change could also cause a decrease in underwater grasses, more "dead zones" of low oxygen, more annual precipitation and a resulting increase in the flow of pollution, fewer wintering waterfowl, and a change in the types of plants and animals that live in the area.

is converted to the graph of 18 facts shown in Figure-1. Grey boxes starting with r: are representing roles and blue boxes starting with e: are events. Each box is represented by a lemma but in fact represents a concept in de the underlying data. The numbers after each lemma indicate the confidence score for each concept. You can for example see that *pollution* occurs twice with different scores and different roles. The type of role is indicated by the colour and label of the edges. The boxes for *pollution* thus represent different interpretations of the system. By setting a threshold, the graph can be reduced facts with higher confidence scores.

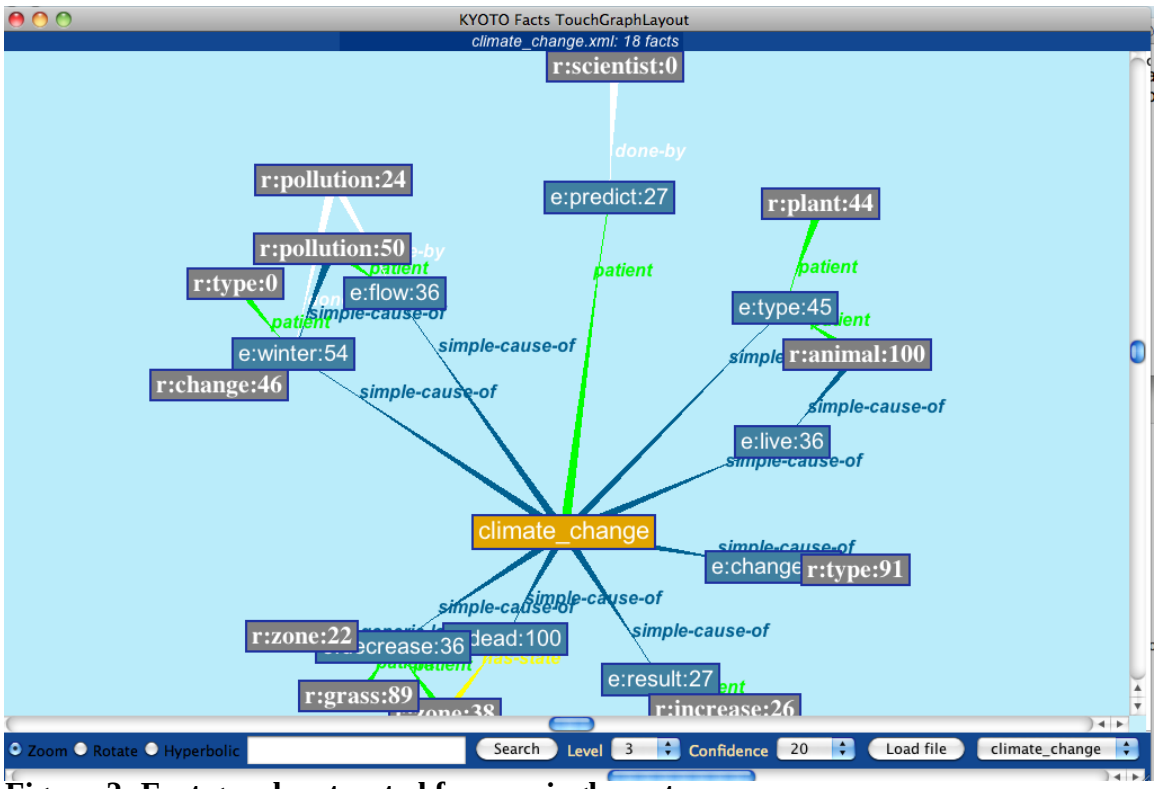

Figure-2: Fact graph extracted from a single sentence

The first year of the project resulted in a detailed specification of the type of information and knowledge that is involved in the environment domain and the way in which the users need to handle the knowledge and information. This was discussed and presented in the 1st project workshop in February 2009. In the first year, we also completed the design phase of the project which resulted in the specification of various representation formats that are compliant to current standards. Finally, we developed first releases for all modules in KYOTO and processed databases in all languages, including a first mock-up version of semantic search for simulated facts.

At the end of the second year of the project, we released a first version of the integrated system. This system can apply all the analysis steps from text to facts. Different installations have been built for the languages in KYOTO that process text in the same way. Using this system, we built a database of facts for 4,625 English documents on estuaries.

In the third and final year of the project we worked on evaluation and further improvement of the system. Four types of evaluations have been carried out:

- Environmentalist used the Wikyoto system for building a domain wordnet from the terms that were extracted from the estuary database;
- We organized and participated in the SemEval2010 task for domain-specific word-sense-disambiguation;
- We developed an evaluation frame work for evaluating the facts mined in KYOTO. This was used to evaluate the KYOTO fact extraction against a manually created gold-standard . It was also used for an open-competition task on event-mining that we organized for the 2nd KYOTO workshop, held in January 2011 in Gifu (Japan).

– We developed an end-user evaluation framework for comparing searches in the KYOTO output with a baseline system and a mash-up system. We organized evaluation sessions with 16 students and another one with 10 specialists to evaluate the systems;

In addition to the evaluation, we worked on improving the knowledge resources that are used for processing. We made the knowledge resources for KYOTO less domain dependent, while the whole system can still be tuned to any domain. In particularly, we extended and restructured the KYOTO ontology, which is the formal background for interpreting semantic relations expressed in text. Furthermore, we developed heuristics for extending the mapping relations from the English Wordnet to the ontology. Currently, all the synsets in the English WordNet have been mapped to the ontology, guaranteeing large coverage of concepts. In combination with the option to automatically derive a term database for any domain, this means that the KYOTO system can now be applied to any domain. The WordNet to ontology mapping system also generated many new role relations. These are important to interpret the roles of entities referred to in text. Again, this adds expressive power to KYOTO as a generic system.

The extraction of events is done using so-called Kybot profiles. Kybot profiles are XML files that specify patterns in the analyzed text. These patterns combine textual, structural and conceptual elements in the text. The latter are implications inserted through the above WordNet to ontology mappings and the extended ontology. We developed 261 profiles for English to capture most of the generic relations expressed in the text. Again, these profiles are not specific to the domain and can be used to extract facts from any domain.

We made a selection of 70 profiles that gave the best results on a benchmark document. We applied these patterns to more than 10,000 documents in English for different databases:

- Database with documents on various topics: estuaries, bird migration, climate change, habitat destruction (4,625 documents), the so-called estuary database that was used in the user-evaluation.
- WWF International website (3,271 PDF documents): wwf.panda.org/about_our_earth/all_publications
- Journal of Environmental Biology (791 PDF documents): www.jeb.co.in/journal_issues
- European Environment Agency (713 PDF documents): www.eea.europa.eu/publications
- Hydrology and Earth System Sciences (1,355 PDF documents) www.hydrol-earth-syst-sci.net/volumes_and_issues.htm

For these documents, we generated millions and millions of facts. The results can be downloaded as raw data from the KYOTO website, or can be searched using the cross-lingual conceptual search system that we developed. The data have also been converted to RDF format. As such they can be published to the linked data project[1] and Semantic Web3.0 applications can be built on top of it. We thus achieved to bridge the gap between unstructured text data and semantic web data.

To illustrate the richness of the data we give here further details on the largest database, which is 4,625 documents and 3,091,842 words in size. The patterns yielded almost 1 million information triplets: 118,255 events with 245,563 involved participants, 317,749 dates, 271,734 place relations and 64,604 mappings to countries. The dates and places are entities mapped to ISO dates and

[1] http://linkeddata.org/

GeoNames locations. The mappings from the text yielded 5,075 unique locations and 1,587 dates that anchor the extracted events. The following relations to participants were extracted:

Relation	**Nr. participants**	**Relation**	**Nr. participants**	**Relation**	**Nr. participants**
destination-of	11,033	part-of	2,464	source-of	5,185
done-by	37,096	patient	131,662	state-of	2,575
generic-location	15,883	purpose-of	8,570	use-of	2,093
has-state	5,278	simple-cause-of	23,724		

We also processed smaller databases in the other KYOTO languages. We even applied the system to other domains:

1. Italian Estuary database v2 (about 300 PDFs/ websites)
2. Dutch Estuary database v2 (about 46 PDFs/ websites)
3. Spanish Estuary database v2 (about 47 PDFs/websites)
4. Basque Estuary database v2 (32 PDFs/websites)
5. English pilot database v3 (about 170 PDFs)
6. Po Estuary Database v1 - Italian (about 170 PDFs)
7. Semeval 2010 Database – English (background, train and test data - about 30 PDFs)
8. Semeval 2010 Database – Dutch (about 30 PDFs)
9. Semeval 2010 Database – Italian (about 30 PDFs)
10. Historical event mining in Dutch: the KYOTO architecture was used to extract historical events from news paper articles and historical documents (among which Wikipedia) much in the same way as event mining in KYOTO. We used a different ontology and re-used the base concepts defined in KYOTO, as well as the Dutch wordnet. About 500 Kybot profiles have been developed in Dutch and applied to 100 documents.
11. Medical event mining in English and Dutch: we created a term database from 10 English and 10 Dutch documents in breast cancer. From the term database, we created medical domain wordnets in Wikyoto.
12. Dutch party programs: we processed 11 Dutch party programs using KYOTO. We extracted term databases for each party program to define the topics they are interested in. Furthermore, the KYOTO annotation tool was used to annotate the party programs in KAF with rich information on political opinions and positions. The KYOTO system is further used for mining subjectivity relations and opinions in the near future.

The overall system and the corresponding resources represent a major tool for mining facts from text in any domain. The generic system can be tuned to a domain using the automatically learned term database which is automatically linked to the generic system. Nevertheless, users of the system can build a domain specific wordnet that represents the specific concepts that matter for them from the term database.

The Wikyoto system has been improved during this year to make this process very easy for domain specialists that have no training in linguistics, language engineering and knowledge engineering. Furthermore, we extended Wikyoto with the option to map concepts in the domain wordnet to the ontology using simple interviews, the so-called TMEKO procedure. Likewise, domain specialists can add the important semantic implications for the selected concepts, which are stored in a formal representation, without having to know or see these formal structures.

Through Wikyoto, the end-users in KYOTO, who are environment specialists, built a domain wordnet that can be exploited for extracting facts in their domain. The domain-wordnet adds more precise implications to the processed text, which will make the above generic patterns and resources more effective for the domain. Likewise, Wikyoto allows domain-customization of the complex KYOTO processing by domain specialists not familiar with the technology.

The output of the massively mined facts can be searched by end-users in the semantic search system. This system generates structured tables for facts that express important causal relations and relate these to the places and dates in which they occur. These structured tables provide efficient knowledge and information which is mapped to time-lines to show trends or to google-maps to show regional coherence. The semantic search system was evaluated by students and domain specialists. Through the semantic search, we have shown that we cannot only extract rich factual data from text but that we can also build applications on top of these data. KYOTO brings text-mining to another level where it starts approximating full text search in comprehensiveness of the data handled while maintaining the depth of state-of-the-art information extraction, as is shown in Figure-3. Our benchmark tests show that the extracted facts represent at least 30% of the word occurrences that are found in a full text index. Many more information is found though in the facts which are not represented in the full text index. This indicates how far KYOTO is in terms of a full text representation.

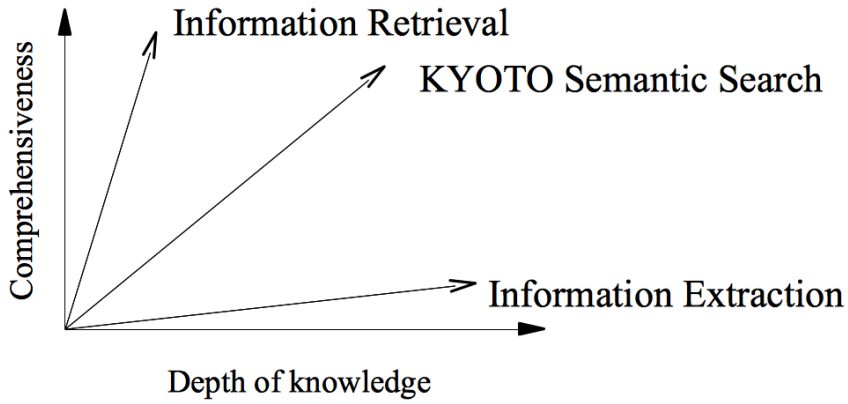

Figure-3: Completeness and depth of information presented

The evaluation of the extracted facts showed that the Kybots extract 40% of the annotated facts with a precision of 50% using optimal settings. The end-user evaluation and the search benchmark showed that the facts represent a large proportion of a full text index and showed no significant decrease in information found. Considering the fact that we are dealing with the first output of the system and that the conceptual representation entails many new opportunities to develop new types of applications and interfaces to information, we firmly believe that our approach is very promising for the future.

In January 2011, we held the 2nd KYOTO workshop in Gifu, Japan to present the results of our system. The workshop was co-sponsored by Toyohashi University of Technology (TUT), Japan (http://www.tut.ac.jp/english/) and The National Institute of Information and Communications Technology (NICT), Japan (http://www.nict.go.jp/index.html). We invited major players in the area of information systems: Google, Yahoo, European Research Council and specialists in the environment domain. We also reserved one day of the workshop for specialists from Japan. At the workshop, we presented the KYOTO system and discussed the results of the user-evaluation using

the search systems that we built. We also presented the results of the open-competition event-mining task that we organized. The workshop has been a successful international event, which has generated many new ideas and plans for future collaboration and innovation along the lines of the KYOTO results.

In the third year we closed KYOTO and KYOTO closed the knowledge cycle: starting from a community tool that allows to connect experts in a domain to share information in textual form, to the modelling of the terms and concepts that occur in these documents across 7 languages, up to the formal representation of this knowledge after being mined from the text, which is presented back to the community to give them a systematic access to the knowledge incorporated in the text.

KYOTO is already exploited in a series of new initiatives and projects of the consortium members, both by technology partners and the users. On the long term, the impact of KYOTO can be enormous. Interpretation of knowledge and information in Natural language is directly managed and maintained by social communities that can organize themselves spontaneously. More importantly, these communities can deploy this knowledge to extract valuable information and facts from their information sources and make this available in new useful ways to the community but also to the outside world. They can do this for and across any language and they are not dependent on knowledge and language engineers.