KYOTO: a wiki for establishing semantic interoperability for knowledge sharing across languages and cultures

Author Name as to Appear in Print
Piek Vossen, VU University Amsterdam, The Netherlands
Eneko Agirre, EHU, San Sebastian, Spain
Francis Bond, Nanyang Technological University, Singapore
Wauter Bosma, VU University Amsterdam, The Netherlands
Axel Herold, BBAW, Berlin, Germany
Amanda Hicks, BBAW, Berlin, Germany
Shu-Kai Hsieh, National Taiwan Normal University, Taiwan
Hitoshi Isahara, NICT, Kyoto, Japan
Chu-Ren Huang, Hong Kong University, China
Kyoko Kanzaki, NICT, Kyoto, Japan
Andrea Marchetti, CNR-IIT, Pisa, Italy
German Rigau, EHU, San Sebastian, Spain
Francesco Ronzano, CNR-IIT, Pisa, Italy
Roxane Segers, VU University Amsterdam, The Netherlands
Maurizio Tesconi, CNR-IIT, Pisa, Italy

ABSTRACT

KYOTO is an Asian-European project developing a community platform for modeling knowledge and finding facts across languages and cultures. The platform operates as a Wiki system that multilingual and multi-cultural communities can use to agree on the meaning of terms in specific domains. The Wiki is fed with terms that are automatically extracted from documents in different languages. The users can modify these terms and relate them across languages. The system generates complex, language-neutral knowledge structures that remain hidden to the user but that can be used to apply open text mining to text collections. The resulting database of facts will be browseable and searchable. Knowledge is shared across cultures by modeling the knowledge across languages. The system is developed for 7 languages and applied to the domain of the environment, but it can easily be extended to other languages and domains.

INTRODUCTION

This chapter describes the KYOTO system for establishing semantic interoperability for text mining and thus for sharing knowledge across languages and cultures. The system can be used by transnational groups in different languages and cultures with the same domain of interest. KYOTO starts from the assumption that language reflects culture and that the linguistic encoding of knowledge and information is therefore culturally biased. Semantic and cultural interoperability is achieved by defining the words and expressions in each language through a shared ontology. An ontology is a formal, language-independent representation of entities that can be used for inferencing and reasoning.

A Wiki environment will help the users to agree on the meaning of the concepts of interest, to share their knowledge and to relate the terms and expressions in their language to this knowledge. This

process is guided by automatic acquisition of terms and meanings from the textual documents provided by the users. The collaborative system will help the users review and edit all acquired information, with a special focus on achieving consensus but also for different views and interpretations across languages and cultures. The users can maintain their knowledge over time and work towards interoperability of terms and language by fine-tuning.

The Wiki environment uses a formal representation for generating knowledge from the conceptual modeling. This representation is language neutral and is not shown to the user directly but can be used by computer software to extract detailed information and facts from a document collection. The extraction process will use the ontological patterns and their relation to the words and expressions in each language so that the information can be interpreted in the same way across these languages and cultures. Likewise, the KYOTO system functions as a cross-lingual and cross-cultural information and knowledge sharing platform.

The system is developed within the KYOTO project (ICT-211423, http://www.kyoto-project.eu/), which is co-funded by the European Union[i] and by (national) funding of Taiwan and Japan. The project started in March 2008 and will end in March 2011. Currently, we completed the specification and design phase and we integrated the first versions of the system components. In the project, we will be working on a restricted set of languages: English, Dutch, Italian, Spanish, Basque, Simplified Mandarin Chinese and Japanese. We will also apply the system to the domain of the environment and specifically to the topic of ecosystem services, a global phenomenon with different linguistic and cultural interpretations. Nevertheless, the system is designed in such a way that it can be used for any language and can be applied to any domain.

The chapter is organized as follows. First, we will describe the situation for the environment domain as a user-case for inter-cultural and cross-lingual information exchange. Next, we will describe the current state-of-the-art in knowledge modeling and information extraction, explaining the short-comings and opportunities. In section 4, we will describe the KYOTO system that we are developing, as a proposal to support the complex knowledge and information modeling in domains such as the environment. Some detailed examples are worked out in section 5, to illustrate the proposed solution.

INFORMATION AND KNOWLEDGE IN THE ENVIRONMENT DOMAIN

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 environment represents a good example of a domain in which inter-cultural and cross-lingual information systems are really required. Experts in the environment domain are under a growing pressure to acquire actual and correct information on very local and unique regions. Different regions in the world share many aspects although each region still is unique in terms of the combination of features. For example, wetlands across the world may share various aspects but each wetland region is also unique as an environmental ecosystem. This makes it, on the one hand, difficult to generalize solutions and present them centrally, but on the other hand it is also clear that knowledge about aspects of each situation can be shared. Due to human development and global changes, these regions change rapidly and environmental experts likewise need to acquire up to date information about the state of nature and environment on a frequent basis.

The domain is also extremely diverse, since it involves many different areas of interest: nature, biology, health, industry, infra-structures, legal aspects, governmental policies, etc. A consequence of the complex and integrated view on ecosystems is that the environment is considered as a service to humanity that is undergoing an enormous pressure due to human activity, with unforeseen consequences both for nature as for humanity. Environmentalists thus use an economic model to describe the dependencies between humanity and nature, where humanity exploits nature as a resource and nature has a certain capacity to deliver these services and to recover from usage. Furthermore, we see many different view points and interests from different cultures and regions. For example, nature in third world countries is related to poverty and economic dependencies. It has a direct value for survival.

We look at information systems for the environment domain from three different angles:

1. What are the types of questions that the environmentalists would like to get answered by an information system?
2. How is this information expressed in the documents and websites that may contain the answers for these questions?
3. How can linguistically and culturally different expressions and views be connected to a unified model of meaning?

To learn more about the first, we conducted a study on the type of questions that experts in the domain would like to ask through information systems. The experts are working in different areas worldwide. This study revealed that the questions are mostly high-level targets that require a lot of knowledge from the domain. Examples of these questions are:

- Which are the most suitable areas in Europe for pro biodiversity business?
- What are the key biodiversity indicators in a certain area?
- What is the effect of hedgerows on air quality?
- What is the impact of dogs on wildlife?
- Are there huge negative effects with regard to eco-networks and alien invasive species?

On the one hand, the questions express abstract causal relations, such as *indicators*, *impact*, *effect*, on the other hand, they contain complex terms such as *pro biodiversity business*, *biodiversity*, *eco-networks*, *alien invasive species*. Any search system will have difficulty matching the abstract relations to specific phrases in the text, i.e. how are causal relations expressed in languages and across cultures. These systems also will have difficulty decomposing complex domain terms such as *pro biodiversity business*, *alien invasive species*. How does the system know that *tourism* and *agriculture* are considered cases of *pro biodiversity businesses*? Similarly, which species are *alien* and *invasive* and when are they considered as such? The answer to these questions is probably different from culture to culture and from region to region. Information systems that allow people to share this knowledge about environmental regions are likely to be hampered by these different views and specific ways of phrasing questions.

Another approach is to look at the language used to talk about environmental issues. This language has interesting features from the perspective of cultural and linguistic encoding. At first sight, one might think that nature and environment across regions and cultures only differ with respect to the types of landscapes, water areas and species that can be found all over the world. However when describing situations in regions, environmentalists use a very rich terminology to refer to the roles of nature in various processes that affect it. In the English documents, *highways* are, for example, called *obstructions for species migration*, and *ecoducts* (bridges over highways between nature areas) are referred to as *connectors* that represent solutions to these obstructions. Another example is the term *corridor*: *hedgerows* are form example called *corridors for wildlife*. This term is further specialized as

migration corridors, *bird migration corridors*, *commuting corridors*, *dispersal corridors*, *terrestrial dispersal corridors*. All these words can be used to refer to specific areas that play a certain role in a process that is relevant for the ecological domain.

If we consider an unrelated language such as Basque, within the same Western culture area, we see similar terms being used, i.e. *migrazio korridore* and *migrazio bide* are equivalent to *migration corridor*. This term is very typical for the Basque country since it is the lowest area in the Pyrenees along which species can more easily migrate. It furthermore has special regions such as wet areas and swamps. In English, these regions are also called *stepping stones* for migration and in Basque they are called *pausaleku*. Such concepts are thus the result of regional circumstances and cultural perspectives.

If we look at Chinese, it is not common to use *corridor* to describe the route taken during migration. Rather, it is directly described by the more general word 路徑 (lu-jing, 'route, course') such as in 遷徙路徑 (chian-shi-lu-jing, 'migration route'), which results in a more abstract term that is more neutral with respect to the protective role. Chinese, on the other hand, provides another interesting case of lexicalization. The basic meaning of the word 環保 is 'protection done in order to prevent environmental damage or pollution', and can roughly be translated with the noun compound *environmental protection* in English. This word can be combined with other words to coin compounds, e.g. 環保團體 (*environmental protection organization* an organization dedicating in environmental protection affairs.), 環保自行車 (*environmental protection bicycle*, environmentally friendly bicycle.), 環保購物袋 (*environmental protection bag*, a re-usable shopping bag) and 環保筷 (*environmental protection chopsticks*, re-usable chopsticks). So the notion of environmental protection is highly lexicalized and can be used productively in combination with many other concepts but it is not combined to form *migration corridor*.

A similar general concept is found in Japanese. The word もったいない */mottainai/* means to 'a sense of regret concerning waste when the intrinsic value of an object or resource is not properly utilized'. This is a very general concept that is actually being proposed in the environment community by the Nobel price winner Wangari Maathai as a generic term for the people's responsibility to the earth.

In Dutch, which is also a compounding language like Chinese, we find other very specific lexicalizations of roles. A good example is represented by the plant species *Urtica* (*Urtica doica* and *Urtica urens*) or nettle in plain English. This plant plays a role in the environment domain in a variety of processes; some of these roles only seem to be lexicalized in Dutch. For instance, nettle serves as an indicator for the amount of nitrogen in the soil. Together with some other plant species that prefer a nitrogen rich environment, these plants are called *stikstofindicator* (nitrogen indicator). At the same time, *Urtica* settles in areas that are for instance influenced by eutrophication, thus suppressing the original vegetation that has already difficulties to survive in the new conditions of the area. If *Urtica* and some other plant species settle in these kinds of areas, Dutch environmentalists refer to them as being a *ruigtesoort*, a plant species that causes (unwanted) rough growth and biodiversity loss. Furthermore, *Urtica* can also have an ecological value as it takes the role of a *waardplant*, a plant species that serves as a kind of host for other organisms. Especially *Urtica* is a *waardplant* for several butterflies that are completely dependant on it for their reproduction, since the caterpillars only feed on the leaves of *Urtica*. Yet another role is that of *pionier* (pioneer) or *pioniersplant* (pioneer plant). This means that Urtica is one of the first plants to settle in former agricultural areas and wastelands, causing better circumstances for other plants and trees to settle. This role is also lexicalized in English. Other more or less domain specific roles that *Urtica* can take are *food, medicine, economic value* (biomass and clothing; in the last role it is lexicalized as *vezelplant* (fiber plant)) and *agricultural value* (used as ecological/natural pesticide).

From these examples, it may be clear that environmentalists use many role-labeling expressions to refer to nature and processes in nature. To sum up: we found obstruction, connector, corridor, stepping stone, sense of regret for the damage to nature, environmental protection role, nitrogen indicator, host plant, nutrient plant, pioneer plant, biomass, fiber plant, ecological/natural pesticide. This is just a

small selection of the many expressions that can be found. These expressions reflect cultural and regional circumstances and can also be applied to certain ranges of specific types of species. So where cultures may agree to some extent on the naming of regions and species, they are very likely going to differ in their usage of these role-labeling terms, either as a reflection of the perspective in the culture or as a result of different lexicalizations across the languages. A culture-aware information system should be able to match these terms to interpretations that can be shared across languages and their respective cultures.

In addition to these roles, we also find culturally specific concepts that differ from culture to culture and that represent truly different things in the world. For example, there are many specific terms across regions for specific water areas or water bodies. In Japanese, we find 天井川 */tenjougawa/* (literally: ceiling river). This is a river that runs with levees built so high that it is running high above the surrounding countryside. The closest in English is 'raised river' or 'raised-bed-river' (a picture can be found in the Japanese wikipedia page http://ja.wikipedia.org/wiki/天井川). In Europe, the term *aqua duct* comes close, which is used both for Roman constructions of water transport and modern versions. Another Japanese example is 溜池 */tameike/*, which is a small reservoir or pond for agricultural use, typically rice paddy irrigation. Yet another example is the Dutch word *wiel* for a small body of water that is only to be found close to dikes. At some point in time the dike gave way, and the force of the water created a pot hole. After the flooding, this pot hole remains in the landscape as small lake. Within the environmental domain, this *wiel* has value as a habitat for birds and fish and is often part of landscape preservation. The English word 'colc' comes close to the notion of *wiel*, but the latter should be regarded as a narrower term. Knowledge about these culture-specific concepts is obviously necessary to be able to share knowledge across cultures and expressed in different languages.

A final more complex example is the division in seasons across cultures and languages. A season is defined in the English lexical database WordNet (Fellbaum 1998) as 'one of the natural periods into which the year is divided by the equinoxes and solstices or atmospheric conditions'. Next it is subdivided into:

- harvest, harvest time (the season for gathering crops)
- haying, haying time (the season for cutting and drying and storing grass as fodder)
- fall, autumn (the season when the leaves fall from the trees) "in the fall of 1973"
- spring, springtime (the season of growth) "the emerging buds were a sure sign of spring"; "he will hold office until the spring of next year"
- summer, summertime (the warmest season of the year; in the northern hemisphere it extends from the summer solstice to the autumnal equinox) "they spent a lazy summer at the shore"
- winter, wintertime (the coldest season of the year; in the northern hemisphere it extends from the winter solstice to the vernal equinox)
- rainy season (one of the two seasons in tropical climates)
- dry season (one of the two seasons in tropical climates)

These seasons map to very different periods of the year across the planet. References to any of these seasons across documents in different languages or even documents in the same language that originate from different parts of the world will be difficult to interpret. It is not enough to know what the equivalences are of *summer* across all the different languages. This has consequences for the interpretation of many climatic data, i.e. measurements of climate properties such as temperature or humidity related to seasons.

To summarize: information systems for the environment domain thus need to be able to:

- handle complex questions for causal relations (e.g. impact, effect);

- between phenomena that are referred to by complex concepts (such as biodiversity business), that are related to certain ranges of specific events (e.g. agriculture, tourism);
- handle role-labeling expressions (e.g. obstructions, stepping stone), that can be applied to ranges of regions and species;
- handle culture specific things such as raised river beds and other water bodies;

We expect that detailed knowledge about such cultural and linguistic differences is required to provide an efficient sharing of knowledge and information.

Knowledge modeling and information systems

Technology development of information systems can be divided roughly in 3 areas:

1. Text based information systems
2. Knowledge mining systems
3. Knowledge repositories

Text based information systems

Text based information systems range from basic statistical indexes to advanced systems that automatically model concepts on the basis of statistical co-occurrence. Many of such systems (commercial and academic) exist for decades now, both on the internet and for intranets. A good overview of current search engines and their characteristics can be found at: http://www.searchengineshowdown.com/. A widely used search engine is Google, which users reported as being their main source of information. Coverage and actuality are two important features of text based information systems such as Google. Through the page-ranking algorithm, Google ensures that the most popular results are delivered first. In addition, Google uses text based matches, preferring results with all matching query words and matches in small distance. More and more, Google also uses techniques to handle linguistic variation, such as fuzzy matching and normalization of inflected words. Still only a very small part of the text on websites is indexed by Google and complex queries, such as the ones discussed above, are handled very poorly.

The major advantage of text based information systems is that they are robust and fast and can handle large amount of data. The disadvantages are:

- They cannot handle ambiguity: a query such as *bats* yields results for baseball, cricket and species;
- They cannot handle different relations: the queries *water pollution*, *polluting water* and *polluted water* will yield either completely different or exactly the same results, depending on the used technology to normalize words. In any case, none of the searches reflect any of these relations properly.
- The user never knows if all results have been found, which text fragments contain the same information or are duplicates;
- The result of a query is a list of text fragments, which cannot be treated as meaningful units that can be used in inferencing or for structuring the information in a useful way (e.g. creating regional maps, or presenting facts on timelines);

The last critique is most relevant here. Because the text is not interpreted, the manipulation of the results is limited. Search engines just list sources ranked for relevance, they do not extract the information and knowledge as such. For example, answering a query for a quantity of species in a region – such as *how many endangered species are there in the Amazon* -- would require that the search engine first determines that Amazon is a region that restricts the relevant species, secondly that

it determines which of these species are currently endangered and thirdly that it cumulates these into a single number. None of this can be done by search engines. Words are matched literally rather than interpreted and the question as such is not understood as being a request for an actual counting.

Knowledge mining systems

Knowledge mining systems do not just build textual indexes but also try to interpret text as meaningful units. They do this using a specific model of the knowledge of interest. Typical text mining applications can for example detect the names of places, people and organizations, or all references to dates. More specifically, they can determine that particular quantities of products are available or have been sold, the stock value of certain assets, the temperature in a specific region, etc. Likewise, they can do a better job of handling questions such as the above.

Peshkin and Pfeffer (Peshkin & Pfeffer 2003) define Information Extraction (IE) as the task of filling template information from previously unseen text which belongs to a predefined domain. Most systems that participated in the Message Understanding Conferences (MUC, 1987-1998) use a pipeline of tools to achieve this, ranging from sophisticated NLP tools (like deep parsing) to shallower text-processing (see for example FASTUS (Appelt 1995)). Currently, the Automatic Content Extraction programme (ACE, http://www.itl.nist.gov/iad/mig//tests/ace) is the main competitive evaluation forum for IE.

Standard IE systems are based on language-specific pattern matching (Kaiser & Miksch 2005), where each pattern consists of a regular expression and an associated mapping from syntactic to logical form. In general, the approaches can be categorized into two groups: (1) the Knowledge Engineering approach, and (2) the learning approach, such as AutoSlog (Appelt et al. 1993), SRV (Freitag 1998), or RAPIER (Califf & R. Mooney 1999). Another important system is GATE (Cunningham et al.2002), which is a platform for creating IE systems. It uses regular expressions, but it can also use ontologies to constrain linguistic patterns semantically. The use of ontologies in IE is a new emerging field (Bontcheva & Wilks 2004): linking text instances with elements belonging to the ontology, instead of consulting flat gazetteers.

IE systems generate structured data from text that can be organized in a useful way, e.g. tables with facts or maps of regions with facts. Furthermore, computer systems can understand the results and take action when required, i.e. send an alert when certain facts have been detected. Another important aspect is that IE results in a single representation of data and not in a list of text occurrences that may express the same or similar data multiple times.

The major disadvantage is that traditional IE systems focus on satisfying precise, narrow, pre-specified requests from small homogeneous corpora (e.g., extract information about terrorist events). Likewise, they are not flexible, are limited to specific types of knowledge and need to be built by knowledge engineers for each specific application. Furthermore, the system needs to know how the knowledge can be expressed in a language. Likewise, most text mining systems are developed for a single domain and a single language. Such systems definitely do not handle knowledge expressed in different languages or expressed and conceptualized differently across cultures. Lately, some promising approaches have been presented for Open Information Extraction (Banko & Etzioni 2008), which scales the relation extraction task to large corpora or the web.

Knowledge repositories

Text mining software detects factual data in text, for example, *the temperature in the last 10 years in the Alps*. Knowledge repositories on the other hand contain more generic knowledge in the form of concepts and relations between concepts. A knowledge repository will make clear that temperature is a physical property, that regions have a temperature and that climates are defined in terms of the average

temperature of a region for a long period of time. Such a generic knowledge repository can be seen as the conceptual model for interpreting relations in text mining.

Knowledge repositories range from weakly/loosely structured data such as thesauri, taxonomies and Wikipedia to formally structured ontologies. Weakly structured knowledge can be used by humans but only to a limited extent by machines. It is also more difficult to merge and combine them, since their meaning is not very explicit. For example, Wikipedia is built by people for people. Humans can read and understand the textual and visual information but computers cannot. The Wiki pages can contain links to other pages that can represent links between concepts, but again not understandable for computers. Ontologies, on the other hand, are founded in logic and the formal representation of concepts. Although they are difficult to read for humans, they can be used by computers to make inferences and reason over knowledge.

A knowledge repository that is special in this respect is WordNet, which is a conceptual knowledge repository based on the English vocabulary. WordNet (Miller 1995, Fellbaum 1998) is a large electronic lexical resource for English, organized as a semantic network (an acyclic graph). It groups words and short phrases into synonym sets: so-called *synsets*; the synsets in turn are interlinked with labeled arcs that represent semantic and lexical relations, such as synonymy, hyponymy (the super-/subordinate relation), meronymy (the part-whole relation), antonym, and entailment relations. As a result, words that are similar in meaning are connected while those whose meanings are unrelated are either unconnected or located in very different parts of the network. WordNet allows one to measure and quantify semantic relatedness, a feature that has made WordNet a popular tool for Natural Language Processing applications that require word sense disambiguation.

Following the English WordNet, similar resources have been built for many other languages and language groups (EuroWordNet, BalkaNet, HindiWordNet, etc.).[ii] Mapping wordnets onto one another reveals cross-linguistic differences in lexicalizations and lexicalization patterns and highlights idiosyncratic aspects of concept-word mappings. Mappings are either directly across languages, or, as in the case of EuroWordNet (Vossen 2004), via a central “interlingua” that serves as the hub for all wordnets. To ensure a language-neutral representation of the concepts underlying the words of each language, the KYOTO project connects the wordnets of its seven languages to a formal ontology, where meanings are represented in a formal, language-independent way (see Section 5).

Unlike a lexicon, which lists the words of a language, an ontology is not bound to language. Rather, it attempts to represent and interrelate concepts that may (or may not) be labeled by a word in one or more language. For example, the root concept in the DOLCE ontology (http://www.loa-cnr.it/DOLCE.html, Masalo et al 2003) is ’particular’; the content of this concept has nothing to do with the common meaning of the word *particular*. A concept may be defined by an axiom in logical form; this underscores its independence from specific lexicalizations and allows for formal operations over concepts in logical form. These logical structures are used for representing the semantic implications of knowledge and are used by machines to make inferences. Like a wordnet, an ontology not only includes, but also structures, concepts into a coherent system by means of relations such as hyponymy (the super-subordinate relation that holds among specific and general concepts) and meronymy (the part-whole relation).

Formal ontologies can be regarded as axiomatized descriptions of categories. Some ontologies model categories as things that exist in the world independent of human conceptualization, see for example the Basic Formal Ontology (http//www.ifomis.org/bfo, Smith 1998). Other ontologies, such as DOLCE, explicitly state that they model knowledge as it is conceptualized given our cognitive and perceptual machinery. Yet another important ontology is SUMO, a Suggested Upper Merged Ontology (Niles & Pease 2001). SUMO is the result of merging and extending various existing ontologies. It is one of the largest public ontologies available and has been mapped fully to the English wordnet and wordnets in other languages: Arabic, Chinese, Dutch, Spanish, Basque.

The different approaches to ontologies have in common that they do not necessarily depend on a particular language or culture. For example, SUMO has been extended to accommodate concepts

originating for the Arabic wordnet (Black et al 2006). In order to ensure clean extensions of ontology hierarchies, Guarino and Welty (2002) therefore proposed the OntoClean method which relies on meta-properties of concepts in an ontology. The meta-property that we focus on is *rigidity*.

Rigid concepts represent properties that are essential to all of their instances, while non-rigid concepts represent properties that exist only contingently for some of their instances. For example, *cat* is a rigid concept but *pet* is not: Each cat must always be a cat under all circumstances or else it ceases to exist. A pet, however, ceases to be a pet when its owner abandons it to the streets or the animal shelter.

Rigidity is an important property for the KYOTO project, and concepts that are represented by nouns in KYOTO's seven languages must be classified in terms of rigidity. Reasoning and inferencing over concepts can only be done accurately when rigid and non-rigid concepts (and the words referring to these concepts) are properly distinguished. For example, a given species might be labeled as *invasive species* in a document. While the species thus labeled (e.g., *kudzu*) will always be that species (i.e., *kudzu* will always be *kudzu* and a type of vine), it may not always be an invasive species (for example, when it is accepted by the native population). An inferencing system therefore must not 'assume' that *kudzu* is an invasive species in the same way that the system 'knows' that it is a vine.

Building ontologies is a difficult and labor-intensive task. Likewise, only a few large and generic ontologies have been built. Moreover, they are built by knowledge experts that are usually not familiar with a specific domain, and they are also usually built for a single language and culture. The harmonization across cultures is achieved top-down. The ontology is built with one language in mind (usually English) and when deployed to other languages, the expressions in these languages simply have to be matched to whatever the ontology dictates. In order to accommodate differences between languages and cultures in a single model of interpretation, we need a model of knowledge representation that defines a shared model for all these languages and cultures and that makes clear how different conceptualizations can be encoded for each. Current knowledge models clearly lack these features.

Combined systems

Some state-of-the-art systems try to combine the above approaches to creating knowledge repositories. Especially in the biomedical domain, systems are being developed that use rich knowledge resources such as bio-medical thesauri together with ontologies to detect data and facts with high precision in large document collections. There are various types of combinations:

1. Mining techniques are used to automatically learn an ontology from text rather than directly extracting facts;
2. Ontologies and other resources are used to support fact mining;
3. Human-crafted databases such as Wikipedia are converted to more formal structures that can be used by computers

The Bootstrep project is a good example of a project that combines these resources *(BootStrep project Web site: http://www.bootstrep.org/bin/view/Extern/WebHome)*. Bootstrep learns the terminology from text and represents the results in an ontology. Then, the terminology and the ontology are used to apply text mining to document collections. Bootstrep is limited to the medical domain. The development of the final term lists and ontology is done by knowledge engineers with the help of medical specialists in the field. However, it is unclear how the medical specialists can maintain the knowledge after the project ends, and it is also unclear how it can be ported to other domains.

Wikipedia is a multicultural and multi-lingual effort that is fully supported by the people themselves. Since its start in 2001, almost 3 million entries have been built for English (date March 2009) and pages have been added for other languages. Many Wiki entries in other languages and

cultures are (partial) translations of English originals, and there is no mechanism to define differences. Unfortunately, the result is not directly usable for computers. DBpedia (http://wiki.dbpedia.org/About) is an initiative to convert the data from Wikipedia into a more structured database. This also has the consequence that knowledge from different language-specific Wikis is merged into a single database model. The latest version of DBpedia (version 3.2) has been provided with a shallow, cross-domain ontology (170 classes and 940 properties). It has been manually created in order to homogenize the representation of all the data mined from Wikipedia info boxes, a particular kind of tabular topic-descriptive template largely adopted in Wikipedia. Despite this, a structured ontological framework for the data mined from Wikipedia that is coherently modeled and showing a global topic representational coverage is still missing in DBpedia. As a consequence, the information gathered is only partially formalized.

Currently there are many ontology editing environments. Among them two important examples are Protégé and OntoWiki. Protégé (http://protege.stanford.edu/, Tudorache & Noy 2007) is an open source platform to edit ontological knowledge collecting and coordinating distinct contributions from different actors with the possibility to define ontology editing workflows and to carry out part of the modifications through the Web Protégé interface. Protégé is actually one of the best-structured ontology editing frameworks, but it is mainly intended for knowledge modeling experts. OntoWiki (http://ontowiki.net/Projects/OntoWiki/, Auer, Dietzhold & Riechert 2006) is a Web-based tool useful to collaboratively edit an ontology and populate it with instances. Even though it has many different knowledge editing facilities, it is still difficult for people who are not trained for this task to correctly achieve a rich and coherent structuring of knowledge.

None of the current combined systems can be generalized and easily deployed to other domains. The technical and scientific nature of for example the bio-medical domain makes it relative easy to detect information in text across languages and cultures since the knowledge is already highly standardized. The interpretation and variation is more limited because the domain is isolated and well-studied. There is a high-degree of consensus at a global level about what terms mean and when they should be used. Such approaches are difficult to transfer to more open domains such as the environment.

The KYOTO system

KYOTO starts with the assumption that the people working in a given field are most qualified to define the meaning of domain terms. They represent a large labor force for encoding and maintaining their own knowledge over time. Furthermore, it is important that these communities be encouraged to define their knowledge so that it can be used for their own benefit to find detailed information. The domain experts should directly see the return of investment of encoding their knowledge. This means that encoding of knowledge should directly lead to more and better knowledge to be extracted from textual sources. However, the domain experts should not be bothered with complex knowledge engineering issues. The process of encoding their knowledge should be as easy as building entries in Wikipedia while the result of this should be as formal as ontological knowledge that can be used by computers to find facts in text.

Another objective of KYOTO is that the definition of knowledge in a community takes place across languages and cultures. Through the Wiki environment, the experts in the field can share their knowledge across languages. This is achieved through a language neutral ontology that will be the backbone for interpreting terms and text. Terms that are acquired for a language are mapped to the ontology and lead to proposals that are seen ‘at the other side’ in another language. Differences and commonalities in conceptualization of the same concepts in the ontology are not only rendered explicit to the users but also resolved in such a way that text from different languages and cultures can be interpreted given the shared backbone. If the system works well, communities will continue to model

their knowledge and achieve consensus on the meaning of the concepts in their domains, as well as what the differences are.

The KYOTO system tries to achieve these objectives in 3 major steps:

1. Terminology and concepts from a domain are automatically acquired for different languages from text collections. This is done by Term Yielding Robots, called Tybots.
2. A Wiki platform, called Wikyoto, allows experts in the field to further define and agree on the meanings of the terms and reach consensus across languages and cultures. The Wikyoto system loads the terms extracted for a language and allows the user to further select the terms and edit the semantic network that is mapped to the shared ontology.
3. A text mining environment uses the terminology and the ontology to extract the relevant information and data from text collections in different languages and generates a single shared repository of data that can be kept up-to-date. This program is called a Kybot, which stands for Knowledge Yielding Robot.

An overview of the system architecture and the involved processes is shown in Figure 1. Documents are uploaded in a shared document base. It is possible to search in this document base using standard text retrieval software. The KYOTO system has specific linguistic processors that apply tokenization, segmentation, morpho-syntactic analysis and some semantic processing to the text in different languages. The semantic processing involves detection of named-entities (persons, organizations, places, time-expressions) and determining the meaning of words in the text using a given wordnet in a language. This process of word-sense-disambiguation is the same for all the languages (Agirre & Soroa 2009, Agirre, Lopez de Lacalle & Soroa 2009). In the current system, there are processors for English, Dutch, Italian, Spanish, Basque, Chinese and Japanese.

The output of this linguistic analysis is stored in an XML annotation format that is the same for all the languages, called the Kyoto Annotation Format (KAF, Bosma et al 2009). This format incorporates standardized proposals for the linguistic annotation of text but represents them in an easy to use layered structure. In this structure, words, terms, constituents and syntactic dependencies are stored separately with references across the structures. All other modules in KYOTO draw their input from these structures.

The process proceeds in 3 cycles. In the 1st cycle, the Tybot will extract the most relevant terms from the documents. The Tybot is a generic program that can do this for all the different languages in much the same way. The terms are organized as a hierarchy with semantic relations and, wherever possible, related to generic semantic databases, i.e. wordnets for each language. The domain experts can view the terms in the term database and edit them, i.e. adding or deleting terms, changing their meaning, adding definitions, changing relations, etc. The result is a domain wordnet in a specific language. Each new term can be seen as a possible proposal to also extend the ontology. Through the ontology, the domain experts can establish the similarities and differences across the languages and hence cultures. These users are called the concept users, since they are involved with the modeling of terms and concepts in their domain.

Whenever a proportion of the domain has been modeled, the output can be used to process further documents. For example, it will be easier to detect occurrences of terms and their meaning when part of the domain has been modeled. This represents the second cycle of the process, which does not involve any human intervention. The result is a collection of documents annotated as KAF that has a richer structure with more precision.

The third cycle of the system involves the actual extraction of data and factual knowledge from the annotated documents by the Kybots. The Kybots use a collection of profiles that represent the type of information of interest. In the profile, conceptual relations are expressed and their realization in a language is achieved through the domain wordnets and so-called expression rules. So-called fact users in the domain can formulate these profiles up front and they can be applied to any document set. They

can create their patterns by selecting examples from the text. They do not need to have any knowledge of the underlying conceptual structure or the linguistic structures. Since the semantics is defined through the ontology, it is possible to detect similar data across documents in different languages, even if expressed differently. The detected data and facts are stored in a factual database, which end-users can browse and search in.

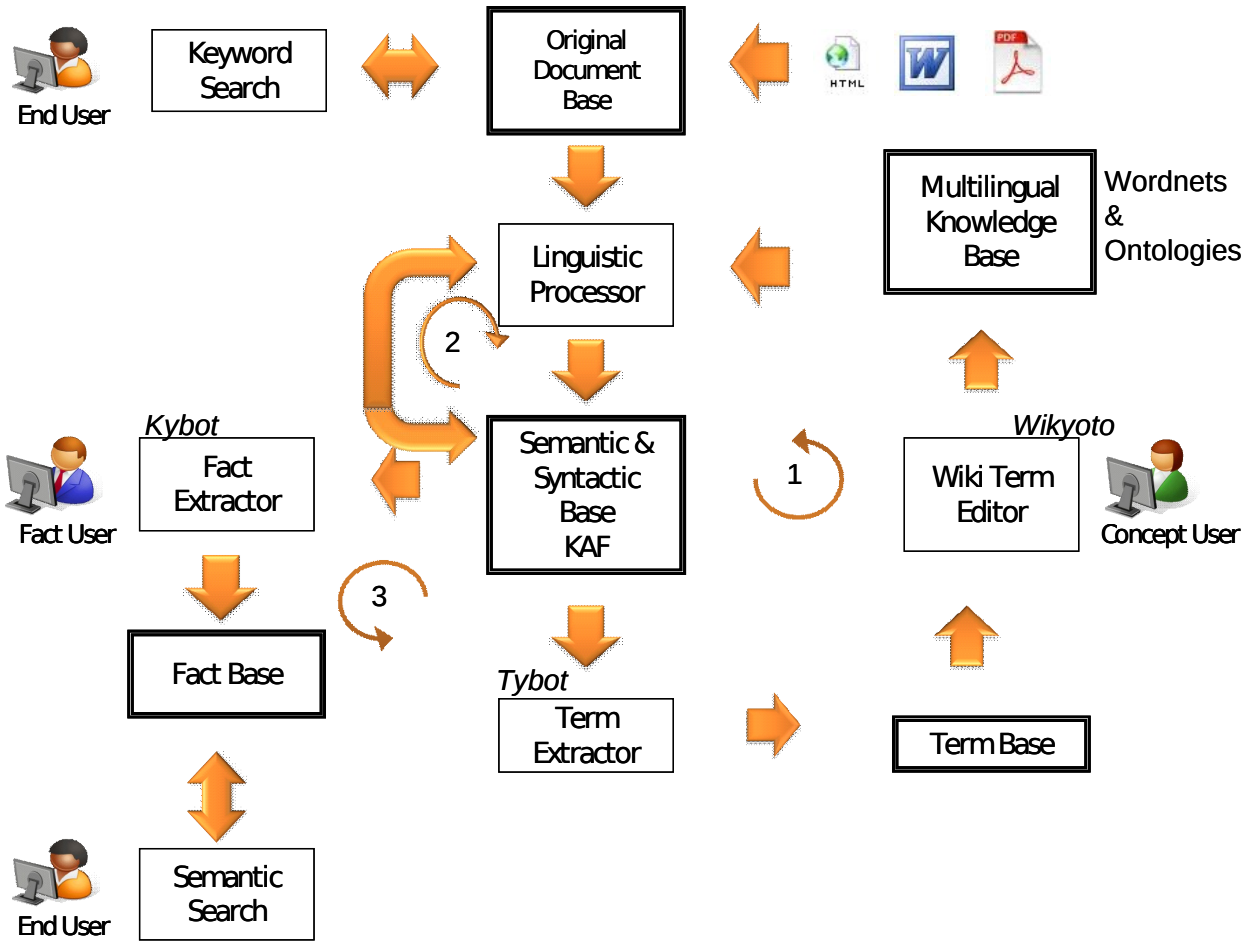

Figure 1. KYOTO system overview.

In the next subsections, we will describe the 3 major modules in more detail.

Tybot

The Tybot detects potential terms (single and multiwords) from the document collections. It takes documents, which are represented in KAF, as input. The extraction of terms is roughly the same for all the languages. The system first generates a maximum list of term candidates using the following structural approach:

1. Any head of a noun phrase is a term: the phrase *the agricultural policy* yields *policy*;
2. Any head of a compound is a term: the Dutch compound *het landbouwbeleid* (the agricultural policy) yields *beleid* (policy) as a term;
3. Any normalized noun phrase is a term: *most agricultural policy* yields *agricultural policy*

4. Any subphrase embedded in a noun phrase that includes the head is a term: *most agricultural policy in the tropics* yields *agricultural policy in tropics*, *policy in tropics*, *agricultural policy*, *tropics*;

Additionally for each multiword term, we extract a parent relation to the head of the phrase from which it is extracted and for each compound a relation to the head of the compound. For example, an *agricultural policy* is a *policy*, indicating that a *agricultural policy* is more specific than *policy*. This gives us a large term hierarchy, in which the top terms are most general, and lower terms are more specific. For about 2,000 English sources on the environment, we extracted over 1 million candidate terms.

Next, we derive a score that indicates the quality of the term, called 'termness', using different features:

1. Whether or not the term occurred independently, or if it was always embedded in a larger term;
2. Salience of a term in the documents based on its frequency count;
3. Number of children in the hierarchy (i.e., more specific terms);
4. Whether or not a term occurs in a defining phrase, where different phrases are used for each language:
    1. X such as Y
    2. X, Y and other Z, X, Y or other Z

The first feature is used to down-rank candidate terms which may not be a valid language unit, preventing over-generation of subphrases. So if *asked question* is always found in the larger phrase *frequently asked question* it is considered a less salient term than if it occurred by itself. The second feature promotes terms that frequently occur in the document collection. Salience of a term can be derived from the hierarchy in several ways, such as plain term frequency, number of documents in which the term occurs, term frequency relative to the frequency in another document collection outside the domain (a reference corpus), the Mutual Information score: co-occurrence frequency of components of a term (e.g., *agriculture* and *policy*) relative to the occurrence frequency of the individual words. Subphrases that are down-ranked can still be kept on the basis of the third feature. If, for example, we also find *rarely asked question*, then *asked question* is promoted since it groups two or more salient terms. The fourth feature provides evidence of salience from the phrases in which the term occurs. The above features are combined in a single confidence score, representing the salience of the term. The confidence score is used to filter the terms in the hierarchy.

If a term is found in the language wordnet, we add the most likely synsets that are detected in the KAF (as the output of the word-sense-disambiguation). On the basis of these synsets, other relations can be added to the term hierarchy from the hyponymy relations in wordnet to group the top terms in the hierarchy. The lesser tops, the more coherent is the term hierarchy, and the richer the tree structure, the more evidence we have for the relevance of terms. For example, isolated words that have no children are less relevant, whereas deep and rich subtrees represent important concepts (expressed by the third feature). Below are examples of term hierarchies for species, extracted from English and Dutch document collections:

English terms related to species

species
- adapted species
- non-native species
- plant species
  - riparian plant species
  - endemic plant species
  - vascular plant species
  - crop plant species
  - indicator plant species
  - domesticated plant species
- animal species
  - endangered animal species
- endangered species
- domesticated species
- taboo species
- bird species
  - threatened bird species
  - breeding bird species
  - widespread bird species
- species in trade
  - species in international trade
- species in zoos
- species in unfavourable population
- species in important habitats

Dutch terms related to species

soort (*species*)
- soorten in agrarische systemen (*species in agricultural systems*)
- karakteristieke soorten van ecosysteemtype (*characteristic species of type of ecosystem*)
- oorspronkelijke soorten (*original species*)
- diersoort (*animal species*)
- kenmerkende soorten (*characterising species*)
- vreemde soorten (*alien species*)

What is salient in one language and not in another can also be seen as an indication of cultural relevance.

Finally, we developed an automatic ontology annotator, called *Rudify* (Herold et al 2009), which can distinguish rigid from non-rigid terms with fairly high accuracy. The tool is based on lexical pattern searches such *as Xs and other Ys, Xs such as Ys* (for rigid concepts*) and X would make a good Y, X stopped being a Y* (for non-rigid concepts). *Rudify* has been evaluated for a set of 215 high-level nominal synsets in the English WordNet, the so-called Base Concepts (Vossen 1998). The accuracy for rigid concepts was 85% and for non-rigid concepts 75%, with a coverage of 57%. The Base Concepts (e.g. *medicine, covering)* are very abstract and general and therefore more difficult to analyze. We expect that the performance will be even better for domain specific words.

The rigidity score of terms can be used to derive the status of the term as a concept for the ontology. Likewise, we can learn that *endangered species* are not a type of species but species in certain circumstances. *Rudify* will be applied to the terms extracted for each language and the scores are combined for the ontology concepts that are associated to terms from different languages. Likewise, we gather cross-linguistic evidence on rigidity in the ontology.

The term hierarchies are the input for the editing process. Using the salience filter, portions of the hierarchy can be shown to the user, who then verifies the relevance of terms as well as relations between the terms.

Kybots

The Kybots try to detect the facts in the text and store the result in the fact database. The Kybot server reads a profile that represents patterns for detecting facts and compiles them into a program that can be applied to any document collection. Kybot profiles consist of three different components: Expression Rules, Semantic Conditions and the Output Template. Once the Kybot profile have been checked and compiled, the resulting Kybot can be applied to the analyzed text (KAF file). Thus, for each analyzed sentence a Kybot is applied using the following rule:

IF (Expression Rules match and Semantic Conditions hold) THEN generate the Output Template

Expression Rules are conditions on the linguistic processing output represented in KAF. They should be flexible enough to deal with the KAF output of all the languages. The Expression Rules represent general morpho-syntactic and semantic conditions on sequences of terms, and relevant pieces of linguistically analyzed text. For instance, the following Expression Rule:

$V=term(@pos="v*" & sense(@sensecode="00151689-v"))

matches in variable $V all occurrences in the text of verbs (@pos="v*") having one particular WordNet sense (@sensecode="00151689-v"), which corresponds to *decrease*, *diminish*, *lessen* or *fall* (in the sense of *decrease in size, extent, or range*). As we are working at a conceptual level this Expression Rule also holds for languages other than English, matching the Spanish verbs *disminuir*, *reducirse*, *consumirse*, *mermar*, *desmoronarse*, or the Italian *decrescere*, *diminuire* or *rimpiccolire*.

Furthermore, the Expression Rules can also encode other Semantic Conditions expressed by resources connected to WordNet, such as Base Concepts (Izquierdo et al. 2007) and Top Concept Ontology (2nd version) (Álvez et al., 2008), WordNet domains (Magnini & Cavaglià, 2000), Suggested Upper Merged Ontology (SUMO) (Niles & Pease, 2001) or DOLCE (Gamgemi et al. 2002).

The conceptual pattern can be the same for different languages but the associated linguistic expressions are unique per language. We expect that a limited number of expressions is needed, which can be combined with an unlimited series of conceptual patterns. Furthermore, we will generate a series of generic profiles, e.g. for relations such as quantities of objects and masses, concentrations of substances in mixtures, time and place expressions, causes, motions, that can be used in any domain.

Finally, domain specific profiles can be added to the collection of patterns using an example-based interface (see Figure 3 below). Fact-users can select text fragments from the document collection that illustrate the type of facts they are interested in. The underlying linguistic and conceptual schema of the example text is used to derive a domain specific profile. Examples of domain specific profiles are: counts of species in regions, decrease/increase of sizes of populations of species, absorptions and emissions of substances, decrease/increase of temperature. The complete collections of profiles or any selection can be deployed to any document collection to mine the facts.

Wikyoto

Wikyoto is the Wiki platform where both domain experts and knowledge engineers can collaboratively browse, refine and enrich all the linguistic and semantic resources exploited in KYOTO. In this way they can maintain and improve the whole system, extending the different kinds of formalized knowledge available in KYOTO and thus making the semantic analysis of data more effective and deeper but also establishing semantic interoperability across languages and therefore cultures.

KYOTO users can interact with Wikyoto through two Web-based environments: the *Wikyoto Knowledge Editor* and the *Kybot Profile Editor*. The concept users use the *Wikyoto Knowledge Editor*

for the creation of domain WordNet extensions concerning all the different languages involved in KYOTO (currently English, Dutch, Italian, Spanish, Basque, Chinese and Japanese) as well as for their shared conceptualization and representation in the KYOTO ontology. Fact users interact with the *Kybot Profile Editor* to identify the generic and domain-specific conceptual patterns and their linguistic expressions, which determine fact extraction by the Kybots. We will analyze both these components of Wikyoto in more details below.

The *Wikyoto Knowledge Editor* allows concept users to browse the Generic WordNets of each involved language and create WordNet domain extensions in a language of their choice, defining new meanings, represented by new synsets and linking them to the ones included in the Generic WordNets. WordNet enrichment is supported by the possibility to navigate the hierarchically organized collections of relevant terms mined by the Tybots. These term hierarchies can be browsed through the *Wikyoto Knowledge Editor* and concept users can create a new synset directly from a particular term. In order to better determine the context of a term, concept users can also visualize all the document occurrences of the same term. These occurrences can be differentiated for the distinct WordNet synsets that are assigned to a term by the word-sense-disambiguation module. Users can either verify and accept these assignments or ignore them. Concept users can also browse a set of relevant SKOS thesauri so as to look for interesting concepts and knowledge structure to exploit as useful suggestions to extend and model Domain WordNets.

In Figure 2, we see a screenshot of the Demo Web Interface of the *Wikyoto Knowledge Editor*, using the English WordNet. On the left side frame there is the 'Static Resources Browser': concept users can browse KYOTO mined term hierarchies, SKOS thesauri and the Generic WordNet of a language of their choice.

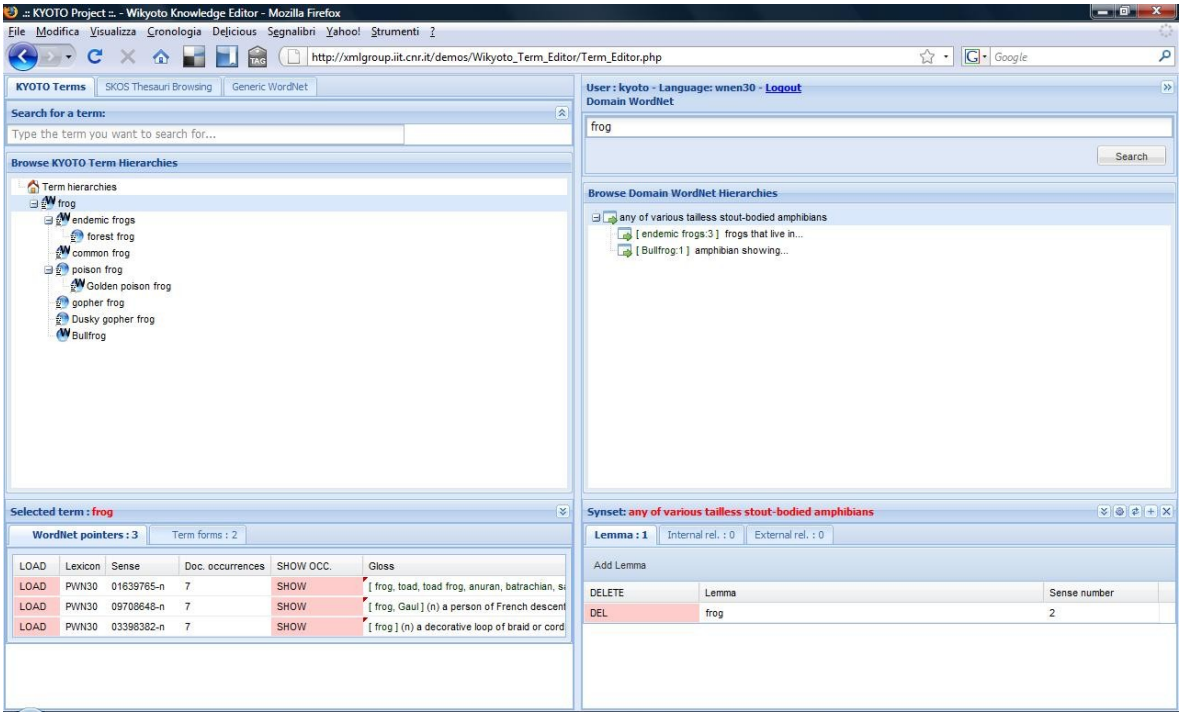

Figure 2. Wiki Term Editor Interface.

In Figure 2 KYOTO term hierarchies are browsed: in particular the term hierarchy related to the term *frog* is visualized, showing many different species of frogs mined from the processed documents. For five of these terms there is some mapping to WordNet synsets as we can notice from the 'W' shown on their left. For instance, the three related synsets are shown for the term *frog* on the lower part of the window. On the right side frame there is the 'Domain Wordnet Browser and Editor', where concept users can create and extend domain WordNets: the domain synset frog as 'Any of various

tailless stout-bodied amphibians' is visualized, along with the hierarchy of all its hyponyms. Concept users can easily drag a term from the left terms hierarchy and drop it over a domain WordNet synset inside the related WordNet hierarchy, thus creating a new hyponym synset.

When a new synset is created, either from scratch or by dragging a term, concept users can directly consult external resources from the *Wikyoto Knowledge Editor* to further define it, i.e. by importing definitions or finding synonyms. In the current version, you can query DBpedia for information or perform simple Google queries. The current demo of the Wiki Term Editor can be accessed through project website: http://www.kyoto-project.eu/.

Besides the editing of a WordNet of a particular language, the *Wikyoto Knowledge Editor* is also used to edit the KYOTO ontology, both to extend it with new domain specific concepts or to map language specific synsets to general ontological concepts. The ontology editor uses the mapping of terms to the Generic WordNet to find the most specific ontology concept that applies to a new synset in the Domain WordNet. Concept users define new ontology concepts when needed as a specification of the language-specific synsets. It is specifically important to define the relations for the role concepts that occur in a language (as discussed in sections 2 and 5). These role concepts need to be related to the processes and properties that matter for the domain and have significant information value. The editing is supported by an analysis of the definition to detect possible relations. Using simple Google patterns of the form "Capitalized plural term+are+plural genus of the term+that", we can find definitions. For example, the Google query "Endangered species are species that" gives the following definitions as the first two hits:

> *Endangered species are species that* face a significant risk of extinction. Such species may be declining in number due to things such as habitat destruction
>
> *Endangered species are species that*, if not protected, are in imminent danger of permanently disappearing from Earth

By marking the most important words, the system can find the relevant processes and properties in the ontology that express these concepts and allow specification of the role-relation for the term. In order to limit the work for the editors, a basic ontology layer is provided that defines the most important processes and properties and relates these to synsets in each of the languages (see section 5 for more details). Since the same ontology is shared by all the languages, the community will be able to more precisely map the domain synsets across languages. They will see what ontological concepts are already defined and mapped in the different languages, and they can judge if this is also lexicalized in their language and whether terms in their language are equivalent. In this way, language independence is obtained and KYOTO's cross-language capabilities can be collaboratively refined and enriched.[iii]

The *Kybot Profile Editor* is the third component of Wikyoto: it can be accessed through a Web interface by fact users in order to collaboratively define and collect the relevant conceptual patterns to be used by Kybots. These conceptual patterns are complex knowledge structures, but they are phrased through natural language examples extracted from KYOTO annotated documents, e.g. *decrease of populations in specific regions and specific periods*. The underlying conceptual patterns for each natural language example are known to the system but as much as possible hidden to the fact users. Fact users have to identify relevant conceptual patterns starting from textual fragments and once identified a conceptual pattern, expert fact users can also refine it by dropping constraints and by associating it to a collection of facts that can be inferred if the same conceptual pattern is found. In this way each pattern defines a type of knowledge through a specific set of constraints, which can be matched against the collection of KYOTO annotated documents to extract the related general facts. A first Demo of the Kybot Profile Editor is available at the KYOTO Project Web Site. Figure 3 shows a screen dump of the current profile editor. We see here the results for a search for the word *decrease*. There are 8 sentences and one of the sentences is selected as an example of the type of fact that the user is interested in.

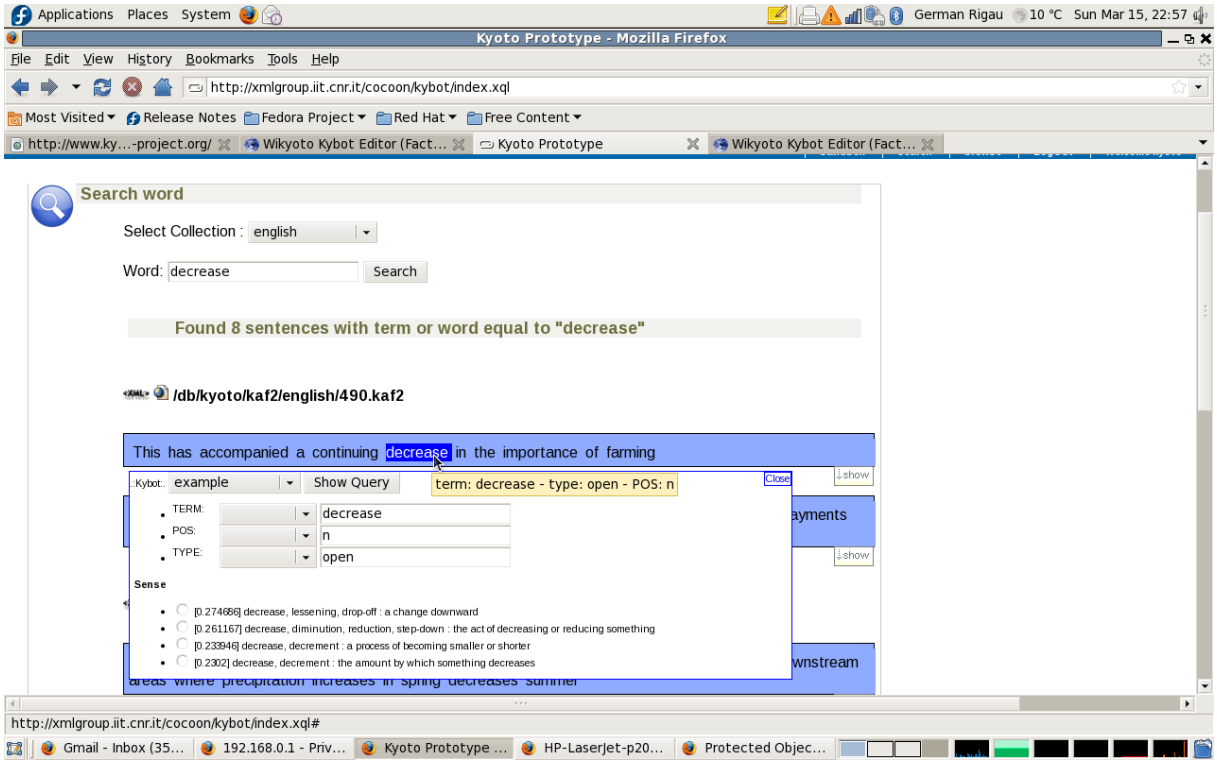

Figure 3. Kybot Profile Editor Interface.

The top window in focus shows the part-of-speech, the semantic type and the different meanings that are related to the occurrence of *decrease* in this sentence. The user can further modify and specify these features, after which a profile is derived from the example. Figure 4 shows the current Sandbox interface in which you can select a Kybot profile and apply it to documents. The profile is here still represented as a complex expression rule, a single matching sentence is shown in the lower box.

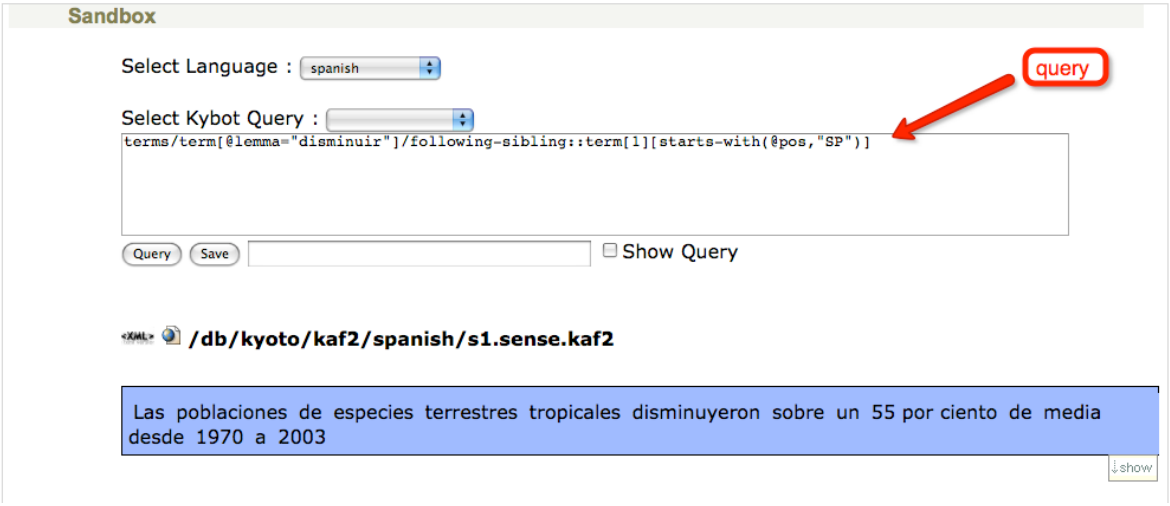

Figure 4. Sandbox for deploying a Kybot Profile.

As explained before, the patterns that are derived as a Kybot profile can be applied across documents from different languages. In that case the morpho-syntactic constraints in the profile need to be dropped or translated to patterns in the target languages.

Encoding cultural specifics in knowledge structures

The major challenge for KYOTO is to establish semantic interoperability across different languages and cultures. We would like to see that KYOTO is used by communities across the world, which operate in the same domain to create a common and shared platform for interpretation of text in different languages. This will reveal cultural differences and similarities. One important layer is the shared KAF representation for documents. Due to its layered structure, it is possible to represent text in different languages in the same way. There is no need for complex linguistic processors. The system will already work if the text is tokenized and constituents with part-of-speech are annotated. This can be done with shallow processors based on rules or machine learning. Certain languages need special modules for word segmentation or morphology, others for multiword recognition and or compounding. The output in KAF is however the same and compatible. Currently, KAF is generated for 7 languages, including non-European languages such as Japanese and Chinese.

More challenging is the interpretation of the terms from the KAF and the way they are mapped to wordnets in each language and the shared ontology. The wordnets in 7 languages are all represented in the same way and related to each other through the English wordnet. These language-specific wordnets and the language-neutral ontology together form a so-called Global Wordnet Grid as defined in Fellbaum and Vossen (1997). Such a grid allows us to define language specific concepts in a language-internal network as well as to anchor it to a neutral ontology. For this we need:

1. A wordnet containing the terms and their meaning in a language/culture;
2. Definitions in the ontology of abstract concepts related to these terms;
3. A definition of concepts that are stable across languages and cultures in the ontology;
4. A definition of how the terms in the wordnet are mapped to the concepts in the ontology;

We will describe this architecture in more detail below.

Following the DOLCE model, the ontology has major hierarchies for endurants (e.g. plants, highways), perdurants (processes such as migration), and qualities (e.g. obstruction, extinction, health). Endurants include both types and roles such as obstruction, migration species, and breeding birds. Events, processes and states are added under the 'perdurant' node in the ontology. Properties are added under the 'quality' node. The following relations are used within the ontology:

- subClassOf, equivalentTo, generic-constituent relations between Endurant:Endurant, Perdurant:Perdurant, Quality:Quality.
- playedBy relation between Role:Endurant.
- hasRole[iv] relation between Perdurant:Role.

For example, the endurants concepts 'plant' and 'animal' have a subClassOf relation to 'organism' and the endurant 'highway' is a subClassOf 'physical-object', The perdurant 'species-migration' is a subclass of 'migration'. The endurants 'migration-role' and 'breeding-role' both have a subClassOf relation to 'species-role' and 'species-role' has a playedBy relation to the endurant species. Finally, a migration-role played by species is part of the perdurant Migration through the hasRole relation.

The current ontology consists of 786 classes. There are layers to the ontology. The basic layer is based on DOLCE (DOLCE-Lite-Plus version 3.9.7) and OntoWordNet. This layer of the ontology has been modified for our purposes (Herold et. al 2009). The second layer consists of concepts coming from the so-called Base Concepts in various wordnets (Vossen 1998, Izquierdo et al 2007). Examples of base concepts are: *building, vehicle, animal, plant, change, move, size, weight*. The Base Concepts are those synsets in WordNet3.0 that have the most relations with other synsets in the wordnet hierarchies and are selected in a way that ensures that each of the more specific concepts is connected to one of the Base Concepts as specific (sub-)hyponyms. This has been completed for the nouns (about

500 synsets) and is currently being carried out on verbs and adjectives in WordNet 3.0. Through the Base Concepts, we will ensure that any synset in the wordnets is mapped to some concept in the ontology either directly or indirectly. The most specific layer of the ontology contains concepts representing species and regions. These concepts were provided by the end users, and in certain cases, concepts have been added to link the domain specific terms to the ontology. This foundational ontology provides the basic building block for the domain experts to add their knowledge.

The ontology is used to model the shared and language-neutral concepts and relations in the domain. Instances are excluded from the ontology. Instances will be detected in the documents and will be mapped to the ontology through instance to ontology relations (see below). There are two relations that we need for this: instanceOf from instances to Endurant, Perdurant, or Quality and instancePlay from instances to Role. Specific entities in discourse, such as an animal called *Donald*, are then instances of a class in the type hierarchy of objects, e.g. Donald *instanceOf* Duck and can play roles, e.g. Donald *instancePlay* BreedingRole. The latter states that Donald could cease being a breeder while the former states that he cannot cease being a duck. Likewise, we will get a clear separation between the ontological model and the instantiation of the model in reality as described in a text.

In addition to the ontology, we will have a wordnet for each language in the domain. The wordnet consists of synsets with synonyms that are lexicalized in each language (partially learned from the domain documents by the Tybots). In addition to the regular synset to synset relations in the wordnet, we will have a specific set of relations for mapping the synsets to the ontology, which are all prefixed with *sc_* standing for synset-to-concept:

1. Synset:Endurant; Synset:Perdurant; Synset:Quality:
    a. sc_equivalenceOf
    b. sc_subclassOf
    c. sc_domainOf
2. Synset: Role
    a. sc_playRole

For each of these relations, the logical implications are defined as follows:

sc_equivalenceOf implies:
- the synset is fully equivalent to the ontology Type
- the synset inherits all properties of the ontology Type
- the synset is Rigid

sc_ subclassOf implies:
- the synset is a proper subclass of the ontology Type
- the synset inherits all properties of the ontology Type
- the synset is Rigid

sc_domainOf implies:
- the synset is not a proper subclass of the ontology Type
- the synset is not disjoint (therefore orthogonal) with other synsets that are mapped to the same Type either through sc_subclassOf or sc_domainOf
- the synset is non-Rigid
- the synset still inherits all properties of the target ontology Type
- the synset is also related to a Role with a sc_playRole relation

sc_playRole implies:
- the synset denotes instances for which the context of the Role applies for some period of time but this is not essential for the existence of the instances, i.e. if the context ceases to exist

then the instances may still exist (see Mizoguchi et al. 2007, for an extension discussion on the semantics of roles).

In this model, we separate the linguistically and culturally specific vocabularies from the shared ontology while using the ontology as a point of interface for the concepts utilized by the various communities.

For the implementation of the model in a domain, we start with the definition of the concepts that are (relatively) stable across cultures and languages. They represent the atomic backbone for interpretation. The species and regions that make up the environment will instantiate concepts in the ontology, e.g., *Uritica dioica* instantiates 'species' and Baltic Sea instantiates 'body-of-water'. Synsets in languages are expected to be equivalent to the concepts 'species' and 'body of water and be related through a sc_equivalenceOf relation.[v]

Clear natural language definitions can be provided for very specific domain concepts, to determine what they are and whether they are the same across languages and cultures. This applies to cases such as Japanese 天井川 */tenjougawa/* (raised river bed) and 溜池 */tameike/* (a small reservoir or pond for agricultural use), and the Dutch *wiel* (water body next to a dike) that we have seen above. Once these rigid concepts are mapped to the ontology, the wordnets of the different languages will provide the language specific words for these concepts. We thus expect that all languages will have words that are equivalent to the concept *Urtica*. On the other hand, only a few languages will have names for *tenjougawa* or *wiel* although they probably have words for more general concepts of bodies of water. Note that the system does not need to have full definitions for all specific concepts, as long as it is indicated to what ontological type they match and what their equivalences are in other languages although a clear definition will help establishing these relations.

In addition, we have seen that environmentalists use many words that refer to the same entities in terms of their roles. To illustrate how these are defined, we will look at a concrete example. Consider the following sentence from the domain text collection on the Humber Estuary in England (UK):

"The highways in the Humber Estuary obstruct the migration of birds."

The relations expressed in this example, need to be modeled by the following types and relations in the ontology or in the wordnet to ontology relations. Because the ontology is still being developed to accommodate perdurants and qualities, this example is intended as a rough sketch:

```
// endurants
(subclass, Road, PhysicalObject)
(subclass, Organism, PhysicalObject)

// roles
(subclass, LocationRole, Role)
(subclass, MigratorRole, Role)
(subclass, MigrationTargetRole, Role)
(subclass, ConstructorRole, Role)
(subclass, ConstructedRole, Role)
(subclass, ObstructingRole, Role)
(subclass, ObstructedRole, Role)
```

```
// perdurants
(subclass, ObstructionPerdurant, Perdurant)
(hasRole, ObstructionPerdurant, ObstructingRole)
(hasRole, ObstructionPerdurant, ObstructedRole)
(playedBy, ObstructingRole, PhysicalObject)
(subclass, MigrationProcess, Process)
(hasRole, MigrationProcess, MigratorRole)
(hasRole, MigrationProcess, MigrationTargetRole)
(playedBy, MigratorRole, Organism)
```

In addition to these basic relations, there can be further definitions of the axioms for these concepts in a formal logical expression.

The language wordnets contain lexemes that can be mapped to any of these elements in the ontology: endurants, perdurants, and roles. Here are some examples:

- {obstruct, obturate, impede, occlude, jam, block, close up}$_{\text{Verb, English}}$
  -> sc_equivalenceOf ObstructionPerdurant
- {obstruction, obstructor, obstructer, impediment, impedimenta}$_{\text{Noun, English}}$
  -> sc_domainOf PhysicalObject
  -> sc_playRole ObstructingRole
- {migration birds}$_{\text{Noun, English}}$
  -> sc_domainOf Bird
  -> sc_playRole MigratorRole
- {migration}$_{\text{Verb, English}}$
  -> sc_ equivalenceOf MigrationProcess
- {migration area}$_{\text{Noun, English}}$
  -> sc_domainOf PhysicalObject
  -> sc_ playRole MigrationTargetRole
- {create, produce, make}$_{\text{Verb, English}}$
  -> sc_ equivalenceOf ConstructionProcess
- {artifact, artefact}$_{\text{Noun, English}}$
  -> sc_domainOf PhysicalObject
  -> sc_playRole ConstructedRole
- {kunststof}$_{\text{Noun, Dutch}}$ // lit. *artifact substance*
  -> sc_domainOf AmountOfMatter
  -> sc_playRole ConstructedRole

Likewise, we represent the general relations involved in the above sentence only once in the ontology and we can relate many terms in the wordnets to a minimal set of ontological elements for the same scenario. The lexicalization of the concepts can differ considerably across languages. As an example, the list of wordnet synsets includes *artifact* in English, which is restricted to objects and *kunststof* (artifact substance) in Dutch which refers to substances.

The ranges of the domain to which role labeling words can refer are typically language-specific. The relation to the ontology clarifies how these different words should be understood and are related to each other. In this respect it is important to realize that the playedBy relation between types and roles in the ontology only encodes a logical constraint, i.e. what is and is not possible. The sc-domainOf relation allows encoding linguistic and culturally specific restrictions on roles. For example, the ontology may express that the FoodRole and PetRole are played by a broad range of types but in languages and cultures these ranges are more specifically restricted and this is reflected in the meaning of their vocabulary. Many animals that may be called pets in Western countries are not considered pets in others; similarly what is called food in China (including dogs and rats) is not considered food in Western countries. English and Chinese will then get different ranges of endurants for the sc_domainOf relation for their synsets for *food.* Our model can exactly accommodate these differences and still make explicit the information that is conveyed by these languages that can be understood across them.

The other major challenge to arrive at semantic interoperability is the detection of facts by the Kybots and the representation of these facts. The Kybot profiles bridge the shared conceptual patterns to linguistic expressions in each language and likewise they produce the instantiation of the ontological relations through instances in the world that are described in the text. The interpretation by the Kybots takes us back from the conceptual structure to the KAF representation of the text. When applying a profile to the text, the Kybot needs to resolve the conceptual constraints. The conceptual constraints are expressed in terms of the ontology or through wordnet synsets. In the latter case, they

can be resolved to the corresponding ontological labels, where the above wordnet-to-ontology mappings are used:

1. hyponymy and meronymy relations from synset to synset ultimately relate a word to an ontological concept that matches the constraint;
2. the domain associated with a synset (or its parents) represents rigid concepts that match the ontological constraint

The above example is then represented in terms of a neutral ontology-instantiation as follows, where instances are represented by numbered variables:

(instanceOf, 0, Location) <!—Humber Estuary →
(instanceOf, 1, Road)
(instanceOf, 2, Organism)
(instanceOf, 3, ObstructionPerdurant)
(instanceOf, 4, MigrationProcess)
(instanceOf, 5, ObstructingRole)
(instanceOf, 6, ObstructedRole)
(instanceOf, 7, MigratorRole)
(instanceOf, 8, LocationRole)
<!—obstruction →
(instanceHasRole, 3, 5) <!—obstruction involves an obstructing role →
(instanceHasRole, 3, 6) <!—obstruction involves an obstructed role →
(instanceHasRole, 3, 8) <!— obstruction takes place in location →
(instancePlay, 1, 5) <!—highways play this obstructing role →
(instancePlay, 2, 6) <!—birds play this obstructed role →
(instancePlay, 0, 8) <!— Humber Estuary plays LocationRole→
<!—migration →
(instanceHasRole, 4, 7) <!— migration involves a migrator role →
(instanceHasRole, 4, 9) <!—migration involves target location →
(instanceHasRole, 4, 10) <!— migration has LocationRole →
(instancePlay, 2, 7) <!—birds play this migrator role →
(instancePlay, 0, 8) <!— Humber Estuary plays location role→

The expressions in the text are now mapped to instances of the ontology concepts through the English wordnet that is connected to the ontology. Furthermore, *obstruction* and *migration* are tied together by the sharing of participants and location within the same sentence. The sentence does not express an explicit causal relation but we can implicitly assume a connection. Note also that not all roles defined in the ontology are also instantiated by the text. These can be assumed to be implied but are not expressed.

We expect a large variation in expressing similar information in text:

- lexicalized roles that imply processes, as they are stored in the domain wordnets;
- explicit references to processes, using verbal lexicalizations, where roles are realized through syntactic subjects, objects or prepositional phrases;
- compounds and multiword expressions that combine roles with processes;
- derivational morphology to refer to roles;

For the above example, the phrase “migration of birds” could also have been phrased as “migration birds”, where the process is implied by reference to the role, or the word obstruction could have been avoided and just the impact can be mentioned:

“The highways in the Humber Estuary have a negative impact on migration birds.”

Our model allows us to represent the implications in the same way, regardless of the way it is expressed in the same language or across languages. We thus learn epistemic instantiations of roles from the document collection in each language. These are stored as language neutral structures, so that they can be shared across cultures and communities.[vi]

When large collections of documents in different languages are processed in this way, we can collect instantiations of the same processes and learn from these data. Through intersecting rigid concepts and roles, we can infer that instances of obstructions are *highways*, *dams*, *canals*, *rivers*, *walls*, *fences*, etc. Similar patterns will be derived for many of the other roles that we have come across: the *Basque country* and other regions are *corridors* and *stepping stones*, *Urtica* is a *pioneer plant* and is also used for making *fibers*, *medicine*, *food*, etc. To some level of detail, these roles need to be represented in the ontology, as well as the processes in which they participate. This allows us to infer how we talk about things in a language in a particular domain. Languages that have terms for these roles are likely to use these terms to refer to instances. For example, *highways* can be referred to by the term *obstruction* in English in the context of species migration: *The increase of obstructions had a dramatic effect on the biodiversity*. The English language can also refer to these roles explicitly, e.g. *The increase of highways obstructs the migration of species, which has a dramatic effect on the biodiversity*. Other languages that do not use the term *obstruction* can only refer explicitly using a verbal phrase if they do this at all.

Collections of such facts are enriched with time and location information that is extracted from the text as well. The result can be organized in useful ways, e.g. all events in the same region within a certain time frame can be grouped together. Representations of all events in a region show the user in a comprehensive way all that happened, providing hints for possible causal associations between events. Furthermore, similar events in a narrow time frame that are derived from different sources can be grouped together or even merged, under the hypothesis that they are likely to refer to the same event. The degree to which they are mentioned in different sources can be seen as evidence for the trustworthiness of the information. Another possibility is to easily find matches of events across different regions scattered over the world that share characteristics. A demonstration of such an information system can be found on the Kyoto website. For such a culturally-aware information system, it is for example possible to systematically compare the information across documents that reside in different cultures and languages. Likewise, we can observe that certain roles and processes are uniquely reported in some and not in others.

CONCLUSIONS

We discussed the specific way in which information is conceptualized by people in the environment domain and how this is realized in the vocabularies of different languages. We noticed that their language is very rich in terms of contextualizing roles and perspectives. Differences in language use make it difficult to share knowledge and information across different linguistic communities. We also provided a solution for solving this in the shared ontology and the way in which the vocabularies are mapped to the ontology. We furthermore described how this knowledge is learned from text collections and can be managed through a Wiki interface by domain experts rather than knowledge experts. These communities can span different cultures and languages but share the same domain. Finally, we explained how the knowledge can be exploited for mining facts from documents in different languages in a uniform way that enables knowledge sharing across cultures.

The KYOTO project is still in progress. Empirical validation of the system and our ideas is on its way. First versions of the various modules have been completed and are being tested on the available document sets for the environment domains. Demos are available on the project website (http://www.kyoto-project.eu/) and integrated and validated results will be made available there soon.

The ultimate test for the KYOTO system is the usage by the people in the community and the effective sharing of information. This is the next goal of the project, where we will deploy it to the environment community and evaluate the capacity to give access to rich semantic data in useful ways.

Acknowledgements

The KYOTO project is co-funded by EU - FP7 ICT Work Programme 2007 under Challenge 4 - Digital libraries and Content, Objective ICT-2007.4.2 (ICT-2007.4.4): Intelligent Content and Semantics (challenge 4.2). The Asian partners from Tapei and Kyoto are funded from national funds.

REFERENCES

Agirre, E., & Soroa, A. (2009) Personalizing PageRank for Word Sense Disambiguation. *Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics* (EACL-2009). Athens, Greece.

Agirre, E., Lopez de Lacalle, O., & Soroa, A. (2009) Knowledge-based WSD and specific domains: performing over supervised WSD. *Proceedings of IJCAI*. Pasadena, USA. http://ixa.si.ehu.es/ukb

Appelt, E. (1995) Sri international fastus system muc-6 results and analysis. In *Proceedings of the Sixth Message Understanding Conference* (MUC-6).

Appelt, E., Hobbs, J., Bear, J., Israel, D., Kameyama, M., & Tyson, M. (1993) Automatically constructing dictionary for information extraction tasks. In *Proceedings of AAAI-93*.

Álvez J., Atserias J., Carrera J., Climent S., Laparra E., Oliver A. and Rigau G. (2008) Complete and Consistent Annotation of WordNet using the Top Concept Ontology. *Proceedings of the 6th international conference on Language Resources and Evaluation*, LREC'08, Marrakesh, Morroco. 2008.

Auer, S., Dietzold, S., & Riechert, T. (2006) OntoWiki - A Tool for Social, Semantic Collaboration. In I. Cruz et al. (Eds.): *Proceedings of 5th International Semantic Web Conference*, Nov 5th-9th, Athens, GA, USA, LNCS 4273, pp. 736.749.

Banko, M. & Etzioni, O. (2008). The Tradeoffs Between Open and Traditional Relation Extraction. In *Proceedings of ACL 2008*.

Black, W., Elkateb. S., Rodriguez, H., Alkhalifa, M., Vossen, P.,Pease, A., Bertran, M., & Fellbaum, C. (2006). The Arabic WordNet Project. In: *Proceedings of the Conference on Lexical Resources in the European Community*. Genoa, Italy.

Bontcheva, K. & Wilks, Y. (2004) Automatic report generation from ontologies: The miakt approach. In *Proceedings of NLDB 2004*.

Bosma, W., Vossen, P., Soroa, A. , Rigau, G., Tesconi, M., Marchetti, A., Monachini, M., & Apiprandi, C. (2009) *KAF: a generic semantic annotation format.* In *Proceedings of the 5th International Conference on Generative Approaches to the Lexicon* Sept 17-19, 2009, Pisa, Italy.

Califf, M. & Mooney, R. (1999) Relational learning of pattern-match rules for information extraction. In *Proceedings of AAAI-99*.

Cunningham, H., Maynard, D., Bontcheva, K., & Tablan, V. (2002) Gate: A framework and graphical development environment for robust nlp tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02), 2002.

Fellbaum C., & Vossen, P. (2007) Connecting the Universal to the Specific: Towards the Global Grid,In: *Proceedings of The First International Workshop on Intercultural Collaboration* (IWIC 2007), Kyoto, Japan, January 25-26, 2007, also in LNCS Vol.4568, Springer-Verlag, 2007.

Fellbaum, C. (Ed.) (1998) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Freitag, D. (1998) Information extraction from html: Application of a general machine learning approach. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 1998.

Gangemi A., Guarino N., Masolo C., Oltramari A., Schneider L. (2002) Sweetening Ontologies with DOLCE. *Proceedings of EKAW*. 2002

Herold, A., & Hicks, A., (2009). Evaluating Ontologies with Rudify Knowledge. *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, Madeira, Portugal, October, 2009.

Herold, A., Hicks, A., Rigau, G., & Laparra, E. (2009) *Kyoto Deliverable D6.2: Central Ontology Version – 1* www.kyoto-project.eu.

Izquierdo R., Suárez A. & Rigau G. Exploring the Automatic Selection of Basic Level Concepts. *Proceedings of the International Conference on Recent Advances on Natural Language Processing* (RANLP'07), Borovetz, Bulgaria. September, 2007.

Kaiser, K., & Miksch, S. (2005) *Information extraction. a survey. In Institute of Software Technology and Interactive Systems*, Vienna, Technical Report Asgaard-TR-2005-6, 2005.

Magnini B. & Cavaglia, G. (2000) Integrating Subject Field Codes into WordNet. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. & Stainhaouer G. (Eds.) *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May- 2 June 2000

Masolo, C., Borgo, S., Gangemi, A., Guarino, N. & Oltramari, A. (2003) *WonderWeb Deliverable D18: Ontology Library*, ISTC-CNR, Trento, Italy.

Miller, G. A. (1995) WordNet: A Lexical Database for English. *Communications of the ACM* 38:39-4.

Mizoguchi R., Sunagawa E., Kozaki K. & Kitamura Y. (2007 A Model of Roles within an Ontology Development Tool: Hozo. *Journal of Applied Ontology*, Vol.2, No.2, 159-179.

Niles, I. & Pease, A. (2001) Formal Ontology in Information Systems. *Proceedings of the international Conference on Formal Ontology in Information Systems* – Vol. 2001 Ogunquit, Maine, USA

Peshkin, L., & A. Pfeffer. Bayesian information extraction network. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI), 2003.

Putnam, H., (1975) The Meaning of "Meaning". In Philosophical papers: Volume 2. Mind, language and reality, Cambridge University Press, 215-271.

Smith, B. (1998) The Basic Tools of Formal Ontology, In Nicola Guarino (Ed.), *Formal Ontology in Information Systems Amsterdam*, Oxford, Tokyo, Washington, DC: IOS Press (Frontiers in Artificial Intelligence and Applications), 1998, 19–28.

Tudorache, T., & Noy, N. (2007) Collaborative Protégé - Stanford Medical Informatics, Stanford, CA 94305, USA - presented at *Social and Collaborative Construction of Structured Knowledge Workshop* at World Wide Web Conference 2007

Vossen, P. (Ed.) (1998) *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.

Vossen, P. (2004) EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an interlingual index. *International Journal of Lexicography* 17.2, 161-173.

Vossen, P., Hicks, A., Rigau, G. Segers, R. (fc) Division of labor for the Global Wordnet Grid in the Kyoto project. *Proceedings of the 5th Global Wordnet Conference*, Mumbai, India, 2010.

KEY TERMS & DEFINITIONS

Semantic interoperability = the degree to which natural language text and resources are anchored to a unified model of meaning across resources and languages

Cultural interoperability = the degree to which knowledge and information is anchored to a unified model of meaning across cultures

Knowledge mining = computer systems that extract knowledge from natural language text

Text mining = computer systems that extract information from natural language text

Information extraction = computer systems that extract information defined in a template from natural language text

Wordnet = lexical semantic database with concepts represented by synonyms in a language, so-called synsets, with semantic relations between these concepts

Synset = set of synonyms that represent a single concept

Ontology = formalized database of conceptual knowledge that can be used by computers to do inferencing

Rigid concept = A concept is rigid if it is essential to all of its instances. For example, the concept *animal* is rigid because everything that is an animal, must be an animal and is an animal for as long as it exists. It cannot cease being animal and change into, for example, a plant.

Role concept = A concept is a role if it is not rigid, which means it is not essential to all or some of its instances. For example, *invasive species* is a role because certain species may become invasive at some point in time and become native at a later point in time.

[i] Co-funded by EU - FP7 ICT Work Programme 2007 under Challenge 4 - Digital libraries and Content, Objective ICT-2007.4.2 (ICT-2007.4.4): Intelligent Content and Semantics (challenge 4.2).

[ii] See http://www.globalwordnet.org for a complete overview of available wordnets for different languages.

[iii] Editors can provide definitions in their native language to map concepts to the ontology but they also need to provide an English definition to explain new concepts in the ontology when created.

[iv] The hasRole relation is compliant to the participant relation in DOLCE. Whereas participant is between perdurant and endurant, hasRole is more specific: between perdurant and role.

[v] Actually, it is not required to add all species and regions to the ontology. There are millions of species and regions and an ontological definition of their semantics is not required for the system to work, as long as they are linked as rigid, and therefore not-disjoint, subclasses of defined ontological types. So, it is already sufficient to know that the synset for *urtica* is related through sc_subclassOf to *Plant* and that it is disjoint to other rigid synsets related to *Plant*. Millions of species can thus reside in the wordnets, while equivalence across languages is indicated through the English wordnet: all language may have synsets that are equivalent to the English synset of *urtica*. This adapted model is called the division of linguistic labor model (Vossen et al fc), along the lines of the principle of the division of linguistic labor by Putnam (1975). The labor is thus divided between the domain experts that can tell what instance of a *Plant* is actually an *urtica*, and the system that only knows what a *Plant* is. Likewise, the ontology can remain relatively small and compact for logical inferencing.

[vi] The information is represented here in an abstract way through the ontology, e.g. *birds* are generalized to *organism*. It is possible to represent data at a more specific level using the wordnet synsets. Knowledge can then still be exchanged and shared across languages if the synsets across these languages match. If that is not the case, the synset hierarchy can be used to find the most specific match across languages.