Knowledge management for more sustainable water systems

Stephen Mounce[1], Christopher Brewster[2], Richard Ashley[1], and Louise Hurley[1]

[1] Pennine Water Group, Department of Civil and Structural Engineering, University of Sheffield, S1 3JD, UK
{s.r.mounce, r.m.ashley, l.hurley,}@sheffield.ac.uk

[2] Department of Computer Science, University of Sheffield, S1 3JD, UK.
c.brewster@dcs.shef.ac.uk

Abstract. The management and sharing of complex data, information and knowledge is a fundamental and growing concern in the Water and other Industries for a variety of reasons. For example, risks and uncertainties associated with climate, and other changes require knowledge to prepare for a range of future scenarios and potential extreme events. Formal ways in which knowledge can be established and managed can help deliver efficiencies on acquisition, structuring and filtering to provide only the essential aspects of the knowledge really needed. Ontologies are a key technology for this knowledge management. The construction of ontologies is a considerable overhead on any knowledge management programme. Hence current computer science research is investigating generating ontologies automatically from documents using text mining and natural language techniques. As an example of this, results from application of the Text2Onto tool to stakeholder documents for a project on sustainable water cycle management in new developments are presented. It is concluded that by adopting ontological representations sooner, rather than later in an analytical process, decision makers will be able to make better use of highly knowledgeable systems containing automated services to ensure that sustainability considerations are included.

Keywords: Water Cycle Management, Ontologies, Text Mining.

1 Introduction

1.1 The Project

The inclusion of careful consideration of water management within the urban planning process is increasingly vital [1]. Reducible and irreducible risks and uncertainties associated with climate, demographic and behavioural changes require preparedness for a range of future scenarios and potential extreme events; to do this

requires the sharing and management of complex information. Communication, knowledge sharing and a common vocabulary across professional domains and between stakeholders is central to more sustainable infrastructure management. New IT methodologies can help structure complex information and facilitate mutual understanding.

The UK Water Cycle management for New Developments (WaND) project (2003-2007) aimed to support the delivery of integrated, sustainable water management for new developments by provision of tools and guidelines for project design, implementation and management. The WaND research consortium comprised work packages researching technical, planning, financial and social science aspects of water management. The tools and procedures developed will assist in co-ordinated decision making across professional boundaries that takes into account risks and uncertainties. The WaND project has provided a case study for application of ontology learning techniques.

1.2 The Water Industry and Ontologies

Ontologies encode knowledge in a domain (a domain being a specific subject area such as the water industry) and also knowledge that spans domains - so called upper level ontologies. Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them and are increasingly valued because of the ever-increasing need for knowledge interchange. Currently, the Web is an incredibly large information source. Yet despite this, the main burden in information access, extraction and interpretation is left to the human user. Tim Berners-Lee coined the vision of a Semantic Web [2] that provides more automated services based on machine processable data and heuristics (methods or strategies, often informal 'rules of thumb' for problem solving) that make use of these meta data. He sees the Web as evolving into a universal medium for data, information, and knowledge exchange.

The Water Industry and related fields are an appropriate area for the application of Ontologies, since the management of complex data, information and knowledge is a fundamental and growing concern. For example, new sensor systems are providing increasing amounts of data. Sensors may be designed by different companies and often downloaded, processed (if at all) and stored on different computer systems. This results in a body of data from different sources that require intelligent algorithms to turn the data into useful information and knowledge for decision makers. Analysis of hydraulic data from water distribution systems for leak detection is one area requiring more sophisticated automated analysis than by simple manual inspection [3].

A variety of tools are available and under constant development for synthesising data into useful information and knowledge to advise decision makers. These tools can include handbooks, hydraulic models, GIS databases etc. Some tools and approaches which make particular use of ontological techniques are now described. The FLUMAGIS project is looking at interdisciplinary development of methods and data processing tools in support of the planning and management of river basins, with an emphasis on implementing the European Water Framework Directive (WFD). Software components enable planners and planning affected citizens to investigate,

debate and evaluate planning measures in a co-operative process. The software tool being developed combines an object ontology (developed in Protégé [4]) with an inference machine with Causal Network (for reasoning about cause and effect) and Petri Nets (for modelling decision processes).

OntoWEDSS (Ontology-based Wastewater Environmental Decision Support System) is a DSS for wastewater management which augments rule-based reasoning and case-based reasoning with a domain ontology. It makes use of the generic WAWO (Waste Water Ontology) ontology and is implemented in LISP [5]. Schwering and Hart [6] look at a case study for the Semantic Translation of the WFD and a Topographic Database with a particular emphasis on semantic translation and portability of Ontologies with the Ontologies in question being the WFD ontology and the topographic ontology.

HarmoniQuA (Harmonising Quality Assurance in model based catchment and river basin management) forms part of the CATCHMOD cluster of European projects; supporting the implementation of the WFD. The HarmoniQuA project uses ontological knowledge engineering techniques to structure the knowledge and make it easily accessible. From this, software-based modelling Knowledge Base and support tools have been developed [7]. The knowledge Base for the tools is built with ontological knowledge engineering techniques and is implemented with Protégé-2000.

The European Project CD4WC (Cost effective Development of Urban Wastewater System for WFD compliance) has explored using an ontological knowledge base as a dissemination tool (a software program consisting of an OWL database and user interface) for stakeholders [8].

The joint Defra/EA project (FD2323) investigated the delivery of the improvement of data and knowledge management within flood and coastal erosion risk management (FCERM). One output of this project has been to describe and document the existence of data, information and knowledge applied to the full range of FCERM decisions, and the current roles and responsibilities for collection, management and use. Detailed ontologies in the form of maps have been assembled to represent the domain and particularly the structure of the relevant stakeholders' organisations [9].

Ordnance Survey is developing knowledge modelling methods to systematically construct conceptual and logical ontologies for flood risk management, while considering the requirements for interoperability between geographical and risk analysis information. Ordnance Survey is also collaborating with Oxford Brookes University to develop a freshwater ecology ontology to be used to test for interaction with their topographic ontology [10].

2 Methods

2.1 Constructing Ontologies

A domain expert defines classes to represent concepts in a domain, with slots (or roles) to represent properties and relationships between the concepts. An ontology,

together with a set of instances, constitutes a knowledge base. An instance has properties with particular values. A class can have subclasses (is-a relation) to represent more specific concepts. Similarly, if some concepts share common properties, they may be generalized by creating a superclass, which holds the common properties. As a simple example, a sub-class of the surface water body class is river. The River Thames is an instance of river. The superclass of the surface water body class is the water body class. However, there is no single unique ontology for any domain [11].

A knowledge base consists of an ontology and a set of instances. Protégé and similar editors can be used for the hand construction of taxonomies (a subject-based classification that arranges the terms into a hierarchy). Alternatively, rather than being human-labour oriented, knowledge acquisition can be machine-aided. Automatic knowledge acquisition as a concept is at least as old as AI itself. It is particularly attractive due to the fact that extracting knowledge from a domain expert (via e.g. interviews or questionnaires) is an arduous, labour intensive task (the 'knowledge acquisition bottleneck'). Therefore excavating knowledge from human artefacts (whether minds or texts) is extremely costly [12]. Data driven knowledge acquisition is particularly desirable as text is massively available on the Web.

2.2 Semi-automatic Creation of Ontologies

Current research by computer scientists is exploring how a set of techniques (based on Information Extraction, Information Retrieval (IR) and Natural Language Processing (NLP)) can be applied to texts for automated Knowledge Acquisition since very complex domains are often extensively described by collections of text documents. These techniques include semi-automated construction of taxonomies of concepts and subsequent discovery of relations among concepts. Such technology is expected to be very important given that much knowledge in industry is kept in informal natural language repositories. Text mining is the process of deriving high quality information from text. Most of the work on text mining combines statistical analysis with various levels of linguistic analysis.

The area has its origins in techniques from the 1960s and 70s for Knowledge Acquisition for AI (e.g. Semantic Network extraction). Work on Thesaurus Extraction for Information Retrieval addressing the extraction of keywords, thesauri and controlled vocabularies has also been a major influence. More recent development for Lexical Knowledge extraction in NLP led to systems being developed for Extraction of Semantic Lexicons from corpora (a corpora is a large and structured set of texts) e.g CRYSTAL [13]. Current work focuses on a series of increasingly complex processes for ontology learning from texts referred to as the Ontology Learning Layer Cake [14]. Figure 1 illustrates this and the key stages are briefly covered here. Term extraction is the lowest level of the ontology learning process. Terms express more or less complex semantic units. Term extraction can consist of a number of techniques that determine most relevant phrases as terms:

- Statistical analysis (such as the comparison of frequencies between domain and general corpora – e.g. Constructed wetland will be specific to the Water Management domain, while River will be less specific to the Water

Management domain). Scores are used such as MI (Mutual Information) used in co-occurrence analysis and TFIDF (Term Weighting which creates a normalised word frequency based on the number of times a word appears in a target document compared with a set of documents).

- Linguistic methods – several layers of analysis including tokenisation, Part-of-Speech and semantic tagging, morphological analysis, extraction of patterns (Adjective-Noun, Noun-Noun etc.) and ignoring names (ICE, IBM etc.) and certain adjectives.

The next step is to identify terms that share (some) semantics, i.e., potentially refer to the same concept (synonyms). As an example, SUDS (sustainable drainage systems) – BMPs (Best Management Practices) used in UK and US practice respectively. These may be multilingual or cross-cultural. Classification (using existing class systems such as WordNet [15]) and clustering (according to similar distributions) are used for this phase. Terms are then candidates for concepts. A concept can be described formally by, for example, a set of instances that the definition of this concept describes. Named-entity recognition and information extraction can then be used to extract instances for a concept from text (e.g. for the River class, extract the River Thames etc.)

Taxonomy extraction produces a taxonomy backbone (is-a relations). Techniques such as lexico-syntactic patterns, distributional similarity and hierarchical clustering (e.g. using Formal Concept Analysis (FCA)) are used. The linguistic paradigms are based on Harris' distributional hypothesis [16] that words that occur in the same contexts tend to have similar meanings: "You shall know a word by the company it keeps" [17]. Relation extraction from text, other than the is-a relation discussed above, has been addressed primarily within the biomedical field due to the ready availability of large numbers of texts. These relations include Part-Of and Attributes (the X of Y). The final level of the Layer Cake is Rules and Axioms i.e. generating logical rules between concepts, but this is an area in its infancy.

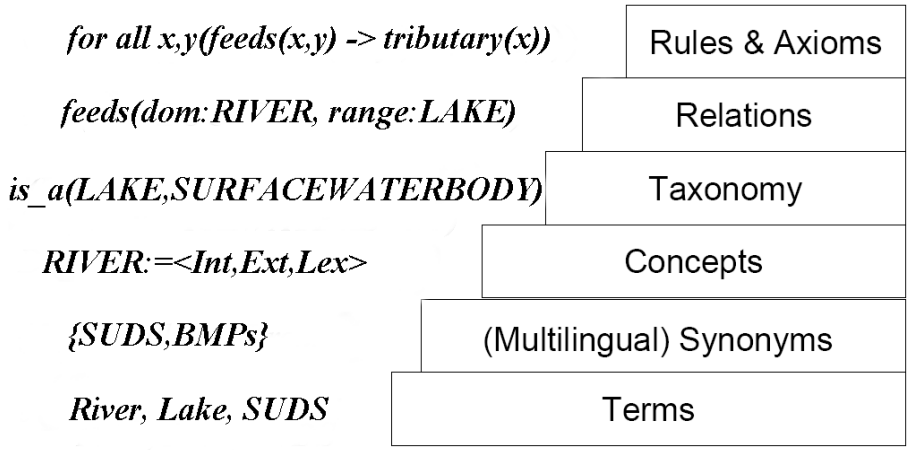

Fig. 1. The Ontology Layer Cake

2.3 Text2Onto

A tool/API called Text2Onto [18] which implements some of the techniques described in section 2.2 was identified as suitable for application to water domain corpora. Text2Onto combines machine learning approaches with basic linguistic processing and incorporates the general purpose ontology WordNet and the GATE (General Architecture for Text Engineering) [19] suite of tools created by the Sheffield University NLP group. Text2Onto introduces two new innovations for ontology learning. Firstly, Probabilistic Ontology Models (POMs) represent the learned knowledge at a meta-level in the form of instantiated modelling primitives which are based on Gruber's Frame Ontology [20] and include concepts (CLASS), the inheritance of classes (SUB_CLASS) etc. By being independent from any particular language, ontology writers can be used to translate to particular representation formats. The POMs also attach a probability to the results learned from the system allowing ranking according to certainty or only showing results above a certain confidence threshold. Secondly, the system monitors for changes in the corpus (documented information) and only recalculates based on those changes, thus avoiding processing the whole corpus every time it changes and enabling the user to trace the evolution of the ontology.

The aim of using software such as Text2Onto is to (semi) automatically generate an ontology. Why do we wish to create such an ontology? Firstly, it makes domain assumptions explicit, providing a community reference point for applications and allows a sharing of a consistent understanding of what information means. Knowledge management is one of the main areas for ontology use. Creating a shared vocabulary (concepts, relations, axioms) of the various actors in a Knowledge management information system allows information to be shared in a more precise way. A by-product of this is knowledge conservation, particularly useful in terms of organisational memory: most big companies lose track of their internal and external data, information, and capabilities. Another consequence is it makes it easier to understand and update legacy data. Ontological software can generate systems that can interface to the organizations' legacy systems and use them as data repositories.

In Text2Onto the user specifies a corpus (a collection of text, html or PDF documents) and starts the graphical workflow editor. The editor allows the selection of algorithms (including combinations) for different ontology learning tasks. The corpus is first pre-processed by a natural language processing component (tokenisation and sentence splitting etc.) before being passed to the algorithm controller. Each algorithm starts by detecting changes in the corpus and updating the reference store accordingly. Finally, it returns a set of requests for POM changes to its caller. After the process of ontology extraction is completed, the POM is presented to the user. Generally, user interaction (e.g. adding or removing concepts, instances or relations) will still be needed for transforming the POM into a high-quality ontology. Finally, the POM can be translated to a suitable representational format (e.g. OWL or RDF) with an ontology writer. It can then be used in any number of applications that need to process the content of information instead of just presenting information to humans including semantic search engines, for filtering and in ontology reasoning products.

3 Application in the WaND project

The part of WaND outlined here aimed to provide support to a range of stakeholders in understanding and implementing water systems that were more sustainable through a flexible framework accessible through the 'portal' - either a website or a stand-alone CD. The stakeholders included: water service providers; local authority planners; housing developers and householders. The portal comprises many hundreds of sources of information and software tools and steering a pathway through these aimed specifically at each stakeholder's individual needs was the primary aim of the ontological investigations.

Text2Onto extracts various types of ontology elements from a given corpus of text documents. The developers (Institute AIFB, University of Karlsruhe) advised the authors on the initial configuration and have been interested in how useful the tool is for application in the water field, as most of the existing experimentation has been based on corpora with technical documents or scientific papers taken from the computer science or knowledge management domains. Several corpora of texts were assembled in the WaND case study. The objective was to achieve the representation of domain knowledge from stakeholder texts (e.g. as regards decision making resources, legislation, organisational memory etc.) in a structured form that would assist a wide range of stakeholders. The corpora were processed with a set of algorithms for each modelling primitive (task). Several algorithms can combine their POM changes in order to obtain a more reliable probability for each primitive. Different pre-defined strategies specify the way individual probabilities are combined and in general a strategy to average the results has been adopted. Corpora 1 consists of WaND project work package summaries concerning techniques and options for more sustainable water management in new developments. Figure 2 shows the top ten concepts based on relevance score for the corpus with the most relevant concept being 'development'. Figure 3 provides the Subclass-of relations by highest relevance, after minor pruning. So for example, 'water supply' is a sub-class of 'water management'.

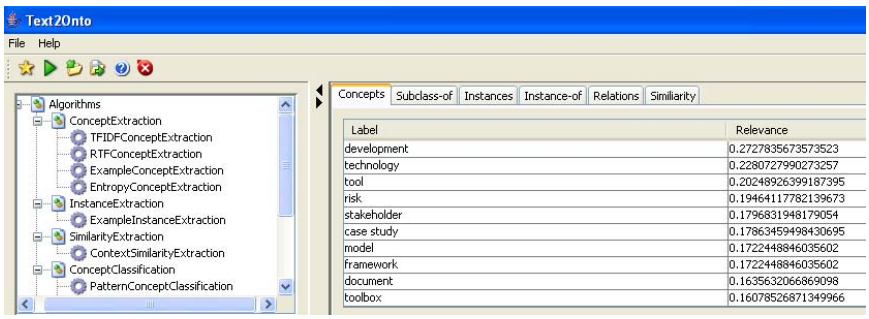

Fig. 2. The ten concepts by relevance for Corpora 1

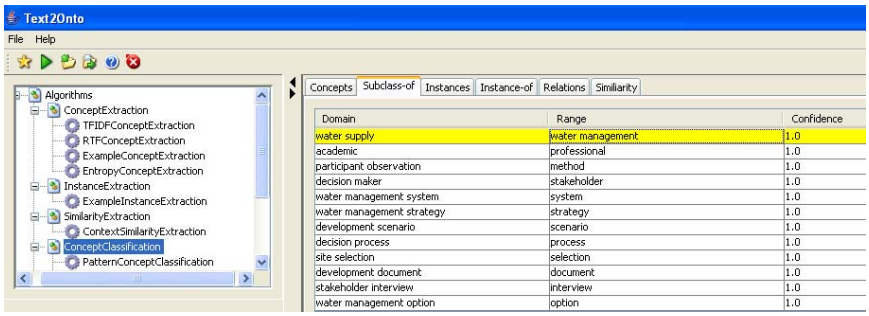

Fig. 3. Subclass-Of by relevance for Corpora 1

Corpora 2 consists of a set of stakeholder texts in this area and included documentation for local authority planners, governmental sustainable development strategy reports, the comprehensive 2006 House of Lords report on water management and interview transcripts on the topic with stakeholders (almost 1000 pages of text in total). The most relevant concept was again 'development'. Figure 4 shows some extracted Instance-Ofs with confidence 1.0. So that 'Breeam' is an instance of 'accreditation scheme' (www.breeam.org).

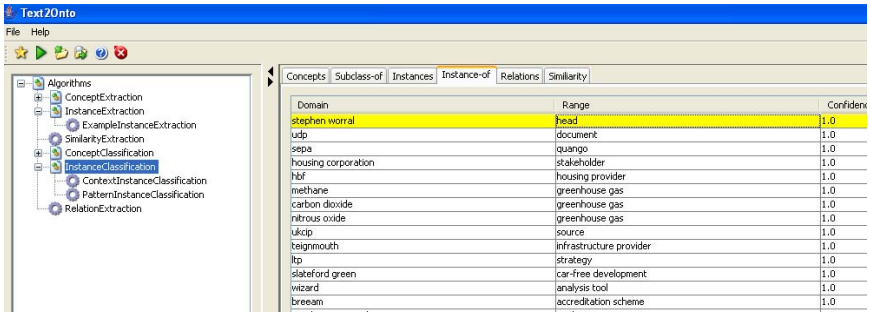

Fig. 4. Instance-of examples for Corpora 2

The extracted primitives for large corpora are, in practice, numerous and some results spurious. Editing is generally required to produce a high-quality ontology. A larger study will need to utilise expert human annotators to evaluate ontologies. Text2Onto was evaluated for the 'knowledge management' domain by five annotators and an average taken of their view on the correctness of the primitives.

In practice, ontologies generated in this way were combined with hand constructed taxonomies to form one aspect of a 'knowledge base' which was used to create the multimedia portal for the WaND project. This portal provided stakeholder decision-making support and incorporates a rudimentary Expert System that illustrates the alternative routes/options open to various stakeholders faced with different stages in the lifecycle of a new housing development, with a focus on more sustainable water management. By traversing a decision tree comprising questions and organized answers, the stakeholder is helped to proceed in the most sustainable way by being

provided with the relevant information, guidance and access to computational and various decision support tools [21]. One way that these decision-making processes of stakeholders can be represented formally is in OWL-DL with a formal reasoner (similar to the biomedical domain).

4 Conclusions

An ontology of a particular domain is not a goal in itself. Developing an ontology is akin to defining a set of data (usually expressed in a logic-based language) and their structure for other programs to use. Ontologies can be seen as meta data that explicitly represent semantics of data in a machine processable way. Problem-solving methods, domain-independent applications such as intelligent search engines and software agents (a piece of software that acts for a user or other program) use ontologies and knowledge bases built from them as data.

In the future, the Semantic Web, constructed from ontologies, will weave together a net linking very large parts of human knowledge and complement it with machine processability. Various automated services will support the human user in achieving goals via accessing and providing information present in a machine-understandable form. This process will ultimately lead to a highly knowledgeable system with various specialized reasoning services that may support human endeavors in nearly all aspects of our daily life. By adopting ontological representations sooner, rather than later, engineers will be able to make full use of these exciting facilities to guide and support their search for knowledge and the way in which this knowledge is used and presented to clients and the wider community.

The work presented here has shown potential for addressing knowledge engineering needs in the sustainable water management domain. It can be seen as providing some first steps in the ultimate aim of developing a comprehensive ontology for this field. This would facilitate greater communication, knowledge sharing and understanding in working towards the complex and challenging aim of sustainable water management.

References

1. House of Lords: Water management Volume 1: report. Science and Technology Committee 8th report of session 2005-6. HL Paper 191-I. The Stationery Office. ISBN 0 10 400871 7 (2006)
2. Berners-Lee, T., Fischetti, M.: Weaving the Web. HarperSanFrancisco, chapter 12. ISBN 9780062515872 (1999)
3. Mounce S.R., Khan A., Wood A.S., Day A.J., Widdop P.D., Machell J.: Sensor-fusion of hydraulic data for burst detection and location in a treated water distribution system, International Journal of Information Fusion, vol. 4, no.3, pp. 217-229 (2003)
4. Noy, F.N., Sintek, M., Decker, S., Crubézy, M., Fergerson, R.W., and Musen, M.A.: Creating Semantic Web contents with Protégé-2000. IEEE Intelligent Systems, vol. 16, no. 2, pp. 60–71 (2001)

5. Ceccaroni, L., Cortés, U. and Sànchez-Marrè, M.: OntoWEDSS: Augmenting Environmental Decision-Support Systems with Ontologies. Environmental Modelling & Software, vol. 19, no. 9, pp. 785-797 (2004)
6. Schwering, A. and Hart, G.: A Case Study for Semantic Translation of the Water Framework Directive and a Topographic Database, Crete University Press. 7th Conference on Geographic Information Science (AGILE), Heraklion, Greece. Seite(n) pp. 503-510 (2004)
7. Scholten, H., Refsgaard, J.C. and Kassahun, A.: Structuring multidisciplinary knowledge for model based water management: the HarmoniQuA approach. Proceedings iEMSs 2004 (2004)
8. Koegst, T., Tranckner, J., Blumensaat, F., Eichhorn, J. and Mayer-Eichberger, V.: On the use of an ontology for the identification of Degrees of Freedom in Urban Wastewater Systems. Proceedings of the 4th International Conference on Water Sensitive Urban Design, Melbourne, Australia (2006)
9. Robinson A D C., Ogunyoye F., Guthrie G., Williams A., Morris J., Romanowicz A., Ashley R M., Cashman A., Blanksby J R.: Improving Data and knowledge Management for Effective Integrated Flood and Coastal Erosion Risk Management. Volume 1. Development of Ontology. Defra/Environment Agency R&D Technical Report FD2323/TR1 (2006)
10. Annoni, A., Bernard, L., Douglas, J., Greenwood, J., Laiz, I., Lloyd, M., Sabeur, Z., Sassen, A.M., Serrano, J.J.: Orchestra: open architecture and spatial data infrastructure for risk management, Proceedings of the International Conference on Risk and Emergency Management, Hannover, Germany, 8-9 June 2005 (2005)
11. Noy, N.F. and McGuiness D.L.: Ontology Development 101: A Guide to Creating Your First Ontology by Noy, N.F. and McGuinness, D.L. SMI technical report SMI-2001-0880, Stanford University (2001)
12. Brewster, C., Iria, J., Ciravegna, F. and Wilks, Y.: The Ontology: Chimaera or Pegasus. Machine Learning for the Semantic Web Dagstuhl Seminar 05071, Dagstuhl, DE (2005)
13. Soderland, W., Fisher, D., Aseltine, J. and Lehnert, W.: CRYSTAL: Inducing a Conceptual Dictionary. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 1314-1319 (1995)
14. Buitelaar, P., Cimiano, P. and Magnini, B.: Ontology Learning from Text: Methods, Evaluation and Applications, Volume 123 Frontiers in Artificial Intelligence and Applications, IOS Press (2005)
15. WordNet, http://www.cogsci.princeton.edu/~wn
16. Harris, Z.: Distributional structure. Word, vol. 10, no. 23, pp. 146-162 (1954)
17. Firth, J. R.: A synopsis of linguistic theory. In J.R. Firth et al. Studies in Linguistic Analysis. Special volume of the Philological Society. Oxford: Blackwell (1957)
18. Cimiano, P. and Voelker, J.: Text2onto - a framework for ontology learning and data-driven change discovery. In Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB'2005) (2005)
19. GATE, http://www.gate.ac.uk.
20. Gruber, T. R.: A Translation Approach to Portable Ontology Specifications, Knowledge Acquisition, 5(2), pp. 1999-2220 (1993)
21. Hurley, L. Mounce, S.R., Ashley, R., Butler, D. and Memon, F.: Flexible Process Frameworks for the Evaluation of Relative Sustainability in Water Management Decisions. International Journal of Technology, Knowledge and Society. Vol. 2, Issue 7, pp. 57-67 (2007)