This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

Excavating Grey Literature: A case study on rich indexing of archaeological documents by the use of Natural Language Processing Techniques and Knowledge Based resources.

Andreas Vlachidis, Ceri Binding, Keith May, Douglas Tudhope

***Abstract*— The paper describes the use of Information Extraction (IE), a Natural Language Processing (NLP) technique to assist 'rich' semantic indexing of diverse archaeological text resources. Such unpublished online documents are often referred to as 'Grey Literature'. Established document indexing techniques are not sufficient to satisfy user information needs that expand beyond the limits of a simple term matching search. The focus of the research is to direct a semantic-aware 'rich' indexing of diverse natural language resources with properties capable of satisfying information retrieval from on-line publications and datasets associated with the Semantic Technologies for Archaeological Resources (STAR) project in the UoG Hypermedia Research Unit.**

The study proposes the use of knowledge resources and conceptual models to assist an Information Extraction process able to provide 'rich' semantic indexing of archaeological documents capable of resolving linguistic ambiguities of indexed terms. CRM CIDOC-EH, a standard core ontology in cultural heritage, and the English Heritage (EH) Thesauri for archaeological concepts are employed to drive the Information Extraction process and to support the aims of a semantic framework in which indexed terms are capable of supporting semantic-aware access to on-line resources. The paper describes the process of semantic indexing of archaeological concepts (periods and finds) in a corpus of 535 grey literature documents using a rule based Information Extraction technique facilitated by the General Architecture of Text Engineering (GATE) toolkit and expressed by Java Annotation Pattern Engine (JAPE) rules. Illustrative examples demonstrate the different stages of the process.

Initial results suggest that the combination of information extraction with knowledge resources and standard core conceptual models is capable of supporting semantic aware and linguistically disambiguate term indexing.

***Index Terms*—Natural Language Processing, Ontology Based Information Extraction, Semantic Annotations, CIDOC Conceptual Reference Model.**

I. Introduction

The Semantic Technologies for Archaeological Resources (STAR)[1] project aims to develop new methods for linking digital archive databases, vocabularies and associated unpublished on-line documents, often referred to as 'Grey Literature'. The project aims to support the considerable efforts of English Heritage (EH) in trying to integrate the data from various archaeological projects and their associated activities, and seeks to exploit the potential of semantic technologies and natural language processing techniques, for enabling complex and semantically defined queries over archaeological digital resources. [1][2]

The STAR project has initially chosen the Raunds Roman Archaeological Database along with Raunds prehistoric data, Raunds environmental sampling data and the Silchester LEAP data to address the aim for semantic integration of diverse data sources. The datasets are held in a number of different systems and formats varying from Delilah legacy coma-delimited ASCII outputs and MS Excel DBF files to MS Access and MySQL data formats. In addition, the differing types and origin of the datasets, indicative of the different stages in the excavation and analysis process which archaeological projects tend to follow, has also been a criterion for the inclusion of the datasets in the project. An archaeological excavation produces a range of data about contexts, plans, photos and text- based reports normally held in a project database. Related activities, such as a geophysical survey of the site, environmental samples and soil samples derived for scientific analysis, the study of human and animal bone remains, also produce associated data that relate to each other in a number of meaningful ways. However the various datasets are currently isolated from each other whilst traditional relational databases queries cannot cope with the complexity of relationships reflected between the different stages of an archaeological excavation eg. "Can you find all the samples with Spelt and seeds from Corn Dryers that were associated with 2nd century contexts and which also contained Barley grains" [3].

To achieve semantic interoperability over diverse information resources and to support complex and semantically defined queries, the STAR project has adopted the English Heritage extension of the CIDOC Conceptual Reference Model (CRM-EH). The adoption of CRM-EH ontology by the project is necessary for expressing the semantics and the complexities of the relationships between

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

data elements, which underline semantically defined user queries as in the example given above. The project has completed a data extraction, mapping and conversion to RDF process, facilitated by an interactive custom mapping and extraction utility. Five datasets have been included in the conversion task; producing a triple store of about 3 million RDF statements. Unique identifiers have been assigned to RDF following the dot delimited notation [*URIprefix].entity.database.table.column.ID]*, providing a consistent convention mechanism for unique naming of entities. Finally a CRM based web service has been implemented to enable search capabilities and browsing of the extracted data over CRM-EH relationships.[2][4]

Archaeological grey literature documents from the OASIS corpus (Online AccesS to the Index of archaeological investigationS) also constitute another valued resource for the aims of the STAR project for enabling access to diverse archaeological resources. Grey literature documents hold information relative to archaeological datasets that have been produced during archaeological excavations and quite frequently summarise sampling data and excavation activities that occurred during and after major archaeological fieldwork. [3][5]

The purpose of ‘excavating grey literature documents' is to produce semantic-aware, rich indices of archaeological texts that comply with the ontological definitions of CRM-EH. To achieve this aim, the research reported in this study explores the potential of Natural Language Processing (NLP) techniques and more specifically the use of Information Extraction for identifying textual representations which are capable of supporting the population of rich semantic indices. The study is directed at the incorporation of knowledge resources (EH thesauri) and the ontological model (CRM-EH) to assist and drive the information extraction process. The adopted information extraction technique is influenced by the notion of Object Based Information Extraction (OBIE), while the potential of semantic-aware 'rich' indices is addressed via the use of semantic annotations. The following paragraphs present the method and the results of an Information Extraction exercise, which aimed to extract and relate textual snippets from grey literature documents with the ontological model CRM-EH. The discussion reveals the method and presents the results of an initial exercise which, identified (and linked to their semantic representations) textual snippets of information relating to two CRM-EH ontological entities *E49.Time Appellation* and *E19. Physical Object*, corresponding to archaeological periods and object finds respectively.

II. METHOD

A. *Excavating Grey Literature Documents*

The current method of excavating grey literature documents is based on the use of Information Extraction techniques which incorporate knowledge resources and ontological references to support a rule based Information Extraction approach, capable of providing semantic annotations that comply with the conceptual reference model CRM-EH. The framework employed to support extraction of textual snippets and assignment of their semantic annotations to documents is GATE (General Architecture for Text Engineering). At the core of the extraction mechanism are JAPE (Java Annotation Pattern Engine) rules which encapsulate the logic arguments of the extraction method and express natural language matching patterns responsible of extracting desirable snippets of information. The method incorporates an ontological reference model for assisting the semantic interoperability of the produced semantic annotations. The level of involvement of the ontological structure in rules formulation describe an Ontology Oriented Information Extraction method, since the exploitation of ontological structure in the extraction process is limited compared to the level of ontology contribution normally achieved by OBIE systems [6]. The method also makes use of the EH Thesauri for Periods and Objects Finds to support term extraction from grey literature documents. Overall 2980 thesaurus terms contributedto an information extraction exercise, which performed over a corpus of 535 grey literature documents and resulted in populating an index of approximately 15.500 individual annotations covering the scope of two ontological entities for archaeological periods and finds.

B. *Information Extraction*

Information Extraction IE, a form of NLP technique, is suggested as enabling richer forms of document indexing. Smeaton recognizes the advances that Named Entity Recognition, can offer in identifying index terms for document representation [7]. Similarly Moens reveals that the idea of using semantic information when building indexing representations is not new, and actually has been expressed by Zellig Haris back in 1959 [8]. Information Extraction is not information retrieval; IE tasks do not involve finding relevant documents from a collection but they are rather specific text analysis tasks aimed at extracting specific information snippets from documents. The fundamentally different role of IE does not compete with IR, on the contrary the potential combination of the two technologies promises the creation of new powerful tools in text processing [7][8][9][10].

C. *Ontology*

Ontologies can be understood as conceptual structures that formally describe a given domain by defining classes and sub-classes of interest and by imposing rules and relationships among them to determine a formal structure of 'things'[11][12]. Ontologies and related Knowledge Organization Systems can be employed to advance the operation of information retrieval systems beyond the limits of words to the level of concepts [13]. Ontological concepts can enrich information retrieval tasks by facilitating rich, semantic information seeking activities, both during query formulation and during retrieval selection. Inferences across diverse sources supported by ontological structures are capable of enhancing information seeking activities and mediating retrieval from heterogeneous data resources [12].

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

D. Semantic Annotations

The term Semantic Annotation refers to specific metadata which are usually generated with respect to a given ontology and are aimed to automate identification of concepts and their relationships in documents [14]. These annotations enrich documents with semantic information, while enabling access and presentation on the basis of a conceptual structure, providing smooth traversal between unstructured text and ontologies. In addition, they can aid information retrieval tasks to make inferences from heterogeneous data sources by exploiting a given ontology and allowing users to search across textual resources for entities and relations instead of words [15]. Ideally the users can search for the term 'Paris' and a semantic annotation mechanism can relate the term with the abstract concept of 'city' and also provide a link to the term 'France' which relates to the abstract concept 'country'. In another case employing a different ontological schema the same term 'Paris' can be related with the concept of 'mythical hero' linked with city of 'Troy' from Homer's epic poem The Iliad. Semantic Annotations carry the critical task to formally annotate textual parts with respect to ontological entities and relations. Such annotations carry the potential to describe indices of semantic attributes which are capable of supporting information retrieval tasks with respect to a given ontology.

There has been a considerable amount of effort dedicated over the last years in the design and development of knowledge management systems capable of supporting semantic interoperable access to information. The range of efforts and the variety of projects is reflected to a number of individual tools and web-portals available today. A clear and straightforward distinction of semantic annotation tools can be made under the condition of their operation, hence defining tools as *automatic* or *manual*.

Manual semantic tools like Annozilla, Mangrove, SMORE, and COHSE assist users in the annotation of HTML, XML and text files, enabling collaborative practices in the assignment of user-defined semantic annotations over content. In the case of SMORE, images and emails can also be annotated whereas in most cases, the tools produce annotations that are compatible with W3C standards formats such as RDF and OWL [16][12]

The automatic semantic annotation tools can be described as *adaptive* which employ Machine Learning (ML) techniques or *rule-based*, which make use of hard-coded linguistic and algorithmic rules, in combination with knowledge base resources to provide, in most cases domain-specific, semantic annotations. Adaptive systems can be *supervised* requiring a set of training data from the user in order to adapt to the domain and to provide annotations relevant to the training set or *unsupervised*, where training data and annotations are produced through a bootstrapping iterative process with little or no intervention from the user.

A major drawback of adaptive techniques is that they require a training set to be annotated by human annotators which, in most cases, is a labour intensive task. Unsupervised adaptive techniques, while not requiring labour intensive human intervention, still when performing in full-automatic mode tend to provide results that are not highly competitive against human performance [15].

Rule-based techniques on the other hand, require no training set of annotations but they require expert domain knowledge and hard-coding skills for the construction of rules and extraction patterns. Last but not least ontologies can be incorporated, both in rule-based and in adaptive tools to enhance system operation and to describe the conceptual arrangements of semantic annotations. Usually such information extraction systems are described as ontology based (OBIE) or ontology oriented (OOIE) depending on the level of ontology engagement [6].

E. GATE Application Environment

GATE is the application environment that enables the information extraction exercise to be performed over a corpus of grey literature documents. Described as an infrastructure for processing human language, GATE is an architecture that provides the framework and the development environment for deploying natural language software components. Offering a rich graphical user interface it provides easy access to *language*, *processing* and *visual* resources that help scientists and developers to produce applications that process human language. At the core of the extraction technique are the JAPE rules which are using regular expressions to recognise textual snippets that conform to particular pattern matching rules, enabling a cascading mechanism of finite state transducers over documents. Moreover, a range of available utilities support additional processing activities such as exporting annotations to XML file structures, manipulation of Gazetteers entries and data management of annotations in relational databases. [10]

III. Prototype Development

The task of identifying and presenting textual representations extracted from a corpus of grey literature documents, constitutes the early form of a semantic indexing effort which has been carried out in three stages. The first stage *pre-processing* selected and prepared the corpus documents for the second stage *extraction-phase* which produced the annotations, while the third *post-processing* stage presented the produced semantic representations of documents in form of hypertext pages. The experiment has managed to annotate a corpus of 535 archaeological documents with semantic attributes connected with two CIDOC-EH ontological entities; *E49.Time Appellation* for archaeological periods such as medieval, prehistoric etc, and *E19.Physical Object* for archaeological objects finds such as axe, flint, wall etc. Initial results are encouraging and reveal the potential of the method in generating semantic annotation metadata with respect to CRM-EH ontological model.

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

A. *Pre-processing.*

During the pre-processing stage, the 535 documents of the corpus collection have been transformed from either pdf or msword files to simple text files encoded with character set Latin-1 (ISO-8859-1). The transformation of files was performed using custom shell scripts and the open source applications *pdf2text* and *antiword* running on Ubuntu (hardy) platform. The newly created text files are lacking any style and presentation to enable optimum execution of JAPE rules. While earlier experiments had noted the ability of GATE to process a wide range of file formats, the performance of rules was noticed to be influenced by space and line break criteria that were not uniform across the corpus collection. The use of plain text files with simple line break statements has been adopted to assist consistent execution of JAPE rules among corpus documents.

B. *Extraction-Phase*

The main stage of the experiment was dedicated to information extraction and semantic annotation and has been developed in GATE using a number of available natural language processing resources, knowledge based resources, user-defined JAPE transducers and the CRM-EH ontological model. Two separate extraction techniques were applied during the experiment; a small scale exercise that introduced the ontological model during the annotation process, and a large scale exercise which incorporated both the CRM-EH ontology and knowledge based resources supporting the construction of complex JAPE rules. A range of language processing resources available from the GATE application environment have been used in both experiments to enable the execution of a cascading pipeline of successive processes targeted to annotate particular textual snippets. The processing resources Tokenizer, Sentence Splitter and Flexible Exporter have been used in both exercises to provide the smallest granules of text (tokens), to define stop words and sentences, and to export the resulted annotations in XML format. The knowledge resources (EH Thesauri) have been transformed to gazetteer lists capable of being processed in the GATE environment, covering approximately three thousand terms and grouped into two types; archaeological periods and archaeological object types (object finds). User defined JAPE rules have also been used in both exercises to extract and to annotate information from documents that conformed to the exercise objectives.

The small scale exercise explored the potential of incorporating an event based ontological model (CRM-EH) in a simple information extraction process, for the annotation of terms of archaeological periods. The Ontogazetteer GATE utility has been employed to map gazetteers terms, originating from the EH Thesaurus of Archaeological Periods, to the CRM-EH entity *E49: Time Appellation*. A simple JAPE rule was used for fetching the gazetteer entries and for producing annotations relevant to the selected ontological entity has been invoked. The annotations were assigned the specific URI (Universal Resource Identifier) of the ontological class to enable semantic interoperability of the annotated terms. This initial investigation suggests that the event based nature of the CRM-EH structure poses challenges for conventional OBIE techniques. Hence further investigation is required for revealing the potential of event-based ontological models in the use of semantic aware language processing techniques.

A large scale exercise aimed at identifying ontological entities in texts but did not make implicit use of the ontological structure per se. Instead the ANNIE Gazetteer utility of GATE has been employed to accommodate the volume of the available EH thesauri terms. Based on their origin, gazetteers terms have been assigned a major type attribute; *Time_Appellation* or *Object_Type,* corresponding to the ontological classes *E49.Time Appellation* and *E19. Physical Object* respectively. In addition, a minor type attribute has been assigned to all gazetteer terms, corresponding to the unique identifier of each individual EH Thesaurus term that is accommodated in the gazetteer resource. Exploitation of major and minor types allowed JAPE rules to annotate textual inputs with respect to ontological and terminological references by assigning to textual representations attributes which correspond to particular ontological entities and thesauri entries.

Several JAPE rules have been constructed during the large scale exercise, expressed in the form of patterns and targeted to particular information extraction cases. A simple negation detection rule has been employed initially to match textual entries that are relevant to the ontological entities but bearing a negative meaning such as 'no prehistoric evidence'. Dedicated rules have been employed for matching textual entries to the two ontological classes for periods and physical objects by exploiting the major and minor types of the gazetteers resource. Extended rules have made use of additional gazetteer terms beyond the scope of EH-Thesauri, expanding the matching capability to phrases like 'earlier Roman period'. Based on the period (E:49) annotations, a set of complex JAPE rules have been produced to annotate compound phrases such as 'from late Roman to early Medieval' and to relate them to the ontological class *E52: Time Span*. Last but not least, rules have successfully annotated textual phrases that included both periods and physical objects as for example the phrase 'Roman Coin' or the phrase 'Burnt flint dating to the late Bronze Age', describing a complex pattern matching mechanism that builds on top of earlier produced annotations .

C. *Post-Processing*

The extraction phase has produced a set of XML files containing both contents and the semantic annotations of grey literature documents that have been created during processing of the corpus collection. The objective of the third stage was to use the resulting XML files for making the semantic annotations available in simple HTML hypertext documents. The server side technology PHP has been employed to handle the annotations from the XML files and to generate the relevant web pages. The resultant pages have been organised under a portal given the name 'Andronikos' which presents the annotations of documents and employs AJAX scripts to link annotations to their semantic definitions. The portal makes use of the DOM XML for processing the XML files and revealing the

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

annotations of documents, while integrates with MySQL database server to store relevant thesauri structures. The Andronikos* portal has been developed to assist the evaluation of the extraction phase by making available the annotations in an easy to follow human readable format and by demonstrating the capability of semantic annotations to link textual representations to their semantic definitions.

*(*http://andronikos.kyklos.co.uk/*, restricted access)

IV. RESULTS

The experiment resulted in the production of approximately 15.500 individual semantic annotations distributed over 535 grey literature documents. Formal evaluation methods and measurements on precision and recall rates have only been applied against a single document where human annotators defined the gold-standards for evaluation. Since this is an early experiment the process of defining the gold-standards to conduct a formal evaluation method is under development. Early evaluation attempts have revealed encouraging results with JAPE rules in some cases outperforming human annotators in recall rates. Competing with a machine is hard when it comes to matching word instances in documents which can be overlooked by humans. On the other hand, human annotators presented better precision rates revealing the ability of humans to comprehend content and to suggest rich and elaborate annotations that are hard to match by a rule based logic.

An early evaluation and visual inspection mechanism has been deployed in the Andronikos web portal. A search engine indexing algorithm provided by the open source FDSE project was deployed in the portal to index the web-pages of the semantic annotations and the full text version pages. The search engine was then used to retrieve results from both indexes to visually inspect their ability to respond common search queries. It is anticipated as the study progresses further that formal evaluation methods will be applied to test the efficiency of the annotation mechanism.

V. DISCUSSION

The experiment has revealed the potential of rule-based information extraction techniques to provide semantic annotations to grey literature documents from the archaeology domain. The use of knowledge resources such as thesauri and conceptual structures such as ontologies evidently can assist the construction of sophisticated rules capable of assigning semantic representations to textual instances. Semantic annotations in the form of XML tags can be manipulated by web applications that make use of server side scripting technologies. This initial experiment represents an early attempt for creating semantic annotations that comply with a given ontological structure. The method has revealed the potential of ontology oriented information extraction techniques in identifying textual parts and linking them to their semantic representations, while revealing a number of issues that relate to the capabilities and limitations of the method.

Future developments should seek to overcome complexities imposed by the event-based model, in order to enable exploitation of the model relations and entities. The rule based mechanisms can be elaborated and assisted by a POS (Part of Speech) tagger. It is planned that the next phase of the experiment will be to incorporate POS inputs in JAPE rules to increase the efficiency of the method and to enable reasoning on the syntactical attributes of natural language text. It will involve the exploitation of the CRM-EH ontological model to advance the experiment to the next phase, widening its scope and including additional ontological entities and more sophisticated rule definitions. Expansion of the experiment towards inclusion of additional knowledge resources in the form of glossaries, thesauri and gazetteers is required to enable the expansion of the method to additional ontological entities. Further utilization of the produced annotations is also much desired to enable contribution of semantic annotations to information retrieval tasks. The Andronikos portal incorporates semantic attributes of annotations to simply display links to semantic definition of terms extracted from grey literature documents. It is within the immediate future plans of the study to investigate the method for transforming the produced XML semantic annotation tags to RDF triple statements. Such RDF resources can be used to introduce the semantic annotations of documents to the semantic retrieval mechanism of the STAR project, which uses the semantic technologies SPARQL and JSON for querying RDF triples [2].

VI. CONCLUSION

Today available semantic technologies promise to close the gap between formal knowledge structures and textual representations enabling new access methods to information [16][17]. Sustainable efforts from the digital archaeology domain have been directed towards enabling semantic interoperability of available digital resources. The provision of semantic annotations to grey literature documents is a challenging task, aimed at enabling access of documents on a semantic - conceptual level. The available language processing technologies make it possible today for scientists and developers to produce software applications capable of revealing the semantic attributes of textual elements and associating them with conceptual structures. The study has attempted to provide semantic annotations to grey literature documents of the archaeology domain, following established information extraction techniques (OOIE - OBIE) and using standard tools (GATE). The initial experiment has revealed that available tools and methods are capable of assisting the process of semantic annotations with promising results. The incorporation of ontologies and knowledge resources (gazetteer, thesauri, glossaries) in a rule-based information extraction technique promises to enable rich semantic indexing of grey literature documents. Additional efforts are required for further exploitation of the technique and adoption of formal evaluation methods for assessing the performance of the method in measurable terms.

This paper is a pre-print version presented at the ISKO UK 2009 conference, 22-23 June, prior to peer review and editing. For published proceedings see special issue of Aslib Proceedings journal.

VII. Acknowledgement

An earlier version of this paper was presented at University of Glamorgan, Faculty of Advanced Technology, Postgraduate Workshop

References

[1] Semantic Technologies for Archaeological Resources (STAR) at http://hypermedia.research.glam.ac.uk/kos/STAR/ [accessed 20 April 2009]

[2] Binding C, Tudhope D, May K. (2008) *Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. Proceedings (ECDL 2008)* 12th European Conference on Research and Advanced Technology for Digita Libraries 280–290

[3] May K, Binding C, Tudhope D. (2008) *A STAR is born: some emerging Semantic Technologies for Archaeological Resources.* Proceedings Computer Applications and Quantitative Methods in Archaeology (CAA2008)

[4] Tudhope D., Binding C., May K. (2008). Semantic interoperability issues from a case study in archaeology. In: Stefanos Kollias & Jill Cousins (eds.), Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop SIEDL 2008, 88–99, associated with 5th European Semantic Web Conference, Tenerife

[5] Online AccesS to the Index of archaeological investigationS (OASIS) at http://www.oasis.ac.uk/ [accessed 20 April 2009]

[6] Bontcheva K, Li Y., Cunningham H, (2007) Hierarchical, Perceptron-like Learning for Ontology Based Information Extraction. Proceedings of the 16th International World Wide Web Conference 777–786

[7] Smeaton A.F., (1997). Information Retrieval: Still Butting Heads with Natural Language Processing? In Alan Smeaton's Online Publications.[Online]Available at: http://www.compapp.dcu.ie/~asmeaton/pubs-list.html [accessed 20 January 2009].

[8] Moens M.F., (2006) *Information Extraction Algorithms and Prospects in a Retrieval Context*. Dordrecht: Springer

[9] Gaizauskas R. & Wilks Y., (1998) Information extraction: beyond document retrieval. Journal of Documentation 54(1) p.70–105

[10] Cunningham H. (2005) *Information Extraction, Automatic*. Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier.

[11] Lee B, Hendler J, Lassila O. (2001) *The Semantic Web*. Scientific American 284(5):28–37

[12] Kiryakov A, Popov B, Terziev I, Manov D, Ognyanoff D (2004) *Semantic annotation, indexing, and retrieval*. Web Semantics: Science, Services and Agents on the World Wide Web 2(1):49–79

[13] Tudhope D., Koch T., Heery R. (2006). Terminology Services and Technology: JISC state of the art review at http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

[14] Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F (2006) *Semantic annotation for knowledge management: Requirements and a survey of the state of the art.* Web Semantics: Science, Services and Agents on the World Wide Web 4(1):14–28

[15] Bontcheva K, Cunningham H, Kiryakov A, Tablan V. (2006) *Semantic Annotation and Human Language Technology*. Semantic Web Technology: Trends and Research in Ontology Based Systems John Wiley and Sons Ltd.

[16] Bontcheva K, Duke T, Glover N, Kings I. (2006) *Semantic Information Access*. In Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems John Wiley and Sons Ltd.

[17] Lee B, Hendler J, Lassila O. (2001) *The Semantic Web*. Scientific American 284(5):28–37