Information Extraction Techniques for the Purposes of Semantic Indexing of Archaeological Resources

Andreas Vlachidis, Ceri Binding, 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 Such unpublished online documents are often referred to as 'Grey Literature'.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

It is said that we live in the Information Society. Today more people than ever before manage an escalating volume of information, which in turn creates an even more complex, and increasing volume of electronic environments aimed to satisfy a wide range of information needs [1][2]. Considering the exponential growth of the Web over the last decades and an ever increasing load of on-line information which has been made available, it wouldn't be an exaggeration to say, that the Web today is the primary source of information for many people, if not for most of us [3].

Information needs on the Web are expressed in words in the form of query terms which are intentionally selected by the users and are expected to occur in the retrieved set of documents result. Based on this fundamental principle, contemporary retrieval systems like Web Search Engines have managed to provide wide access to information while being restrained to operate on the word level. The fundamental argument against the above principle is that not all documents use the same words to refer to the same concept and not all users will use the same words to seek for the same concept. Therefore, contemporary search engines systems are ill-equipped to satisfy information needs that are expressed in a conceptual level beyond the limits of words [4].

Information seeking is a complex human activity that can be studied by various theoretical frameworks but in the everyday 'on-line' practise, users are prompted to satisfy their information needs by simply submitting words in a search box. A comparison study over nine search engine transaction logs revealed that the average number of terms that web users employ to express a single query is 2.2 terms [3]. Considering also that only 2% of the users are using operators to explicate their query and that eight out of ten users simply ignore the results displayed beyond the first page, it is not hard to realise the importance individual words have in query formulation.

The Semantic Web is proposed to add logic to the Web for improving user experience and information seeking activities and so to use rules, to make inferences and to choose courses of action that are defined by the meaning of information and not just by a string of words. It is not proposed to be an alternative or separate web but instead the Semantic Web is envisaged as an extension of the current Web [4]. Supporting access to diverse and distributed collections of information the Semantic Web can enable sharing of information and to provide a coherent view to resources where for example 'zip code' and 'postal code' are defined as resources carrying the same type of information. In addition, dealing with *polysemy;* same word carrying different meaning in different contexts (i.e. 'bar') and *synonymy* different words having the same meaning (i.e. car – automobile) can significantly improve user information seeking activities. It is said that the Semantic Web, when properly designed 'can assist the evolution of human knowledge as a whole' [5].

The Semantic Technologies for Archaeological Resources (STAR) project is aligned to the above views and 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 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. To achieve semantic interoperability over diverse information resources the STAR project adopted the English Heritage extension of the CIDOC Conceptual Reference Model (CRM-EH). The CIDOC CRM is an ISO standard core ontology for cultural heritage information aimed to enable information exchange between heterogeneous resources providing the required semantic definitions and clarifications [6]. In addition, archaeological grey literature documents from the OASIS corpus (Online AccesS to the Index of archaeological investigationS) constitute a 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.

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 study explores the potential of Natural Language Techniques and more specifically the use of Information Extraction for identifying textual representations which are capable to support the population of rich semantic indices. The study is directed in the incorporation of knowledge resources (EH thesauri) and the ontological model (CRM-EH) to assist the information extraction process. The adopted information extraction technique is influenced from the notion of Object Based Information Extraction (OBIE) while the potential of semantic-aware 'rich' indices is addressed with the use of semantic annotations. The following paragraphs present the method and the results of an Information Extraction exercise which aimed to extract and to 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, achieved to identify and to link 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 information snippets and assignment of 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 for extracting desirable snippets of information. The method is incorporating the 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 such that can not be described as OBIE. The method also makes use of the EH Thesauri for Periods and Objects/Finds to support term extraction from texts. Overall 2980 thesaurus terms contributed in an information extraction exercise which performed over a corpus of 535 grey literature documents populating an index of approximately 15.500 individual annotations covering the scope of two ontological entities for archaeological periods and finds.

B.Information Extraction

The potential of Information Extraction IR, a form of NLP technique aimed to extract snippets from documents, is suggested to enable richer forms of document indexing. Smeaton recognizes the advances Named Entity Recognition NER, can have in identifying index terms for document representation [7]. Similarly Marie-Francine 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 (IE) is not information retrieval; IE tasks do not involve finding relevant documents from a collection but they are rather specific text analysis tasks aimed to extract 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' [5][11]. Ontologies can be employed to advance the operation of information retrieval systems beyond the limits of words to the level of concepts. 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 to enhance information seeking activities and to mediate retrieval from heterogeneous data resources [11].

D.Semantic Annotations

Semantic Annotations refers to a specific metadata generation and usage schema aimed to automate identification of concepts and their relationships in documents [12]. 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 [13]. 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 of documents 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.

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 provides architecture, a framework and a development environment for developing and 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 rules, enabling a cascading mechanism of finite state transducers over annotations. Moreover, a range of available utilities support additional processing activities such as exporting the produced annotations to XML file structures, manipulation of Gazetteers entries and data management of the 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 web-page. The experiment has managed to annotate a corpus of 535 archaeological documents with semantic attributes connected with two ontological entities of the CIDOC-EH model; *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 to enable linking between ontological entities and their textual representations.

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 performed using custom shell scripts and the open source applications *pdf2text* and *antiword* running on a Ubuntu (hardy) platform. The newly created text files were lacking any style and presentation to enable optimum execution of JAPE rules. It has been noticed in earlier experiments the ability of GATE in processing a wide range of file formats not only plain text files, but rules' performance were noticed to be influenced from 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 incorporated the ontological model during the annotation process, and a large scale exercise which incorporated knowledge resources in the form of gazetteers supporting the formulation 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 aiming to annotate particular textual evidence. 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 respectively. 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 (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.

A 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 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. The event based nature of the CRM-EH structure offered limited support when JAPE rules attempted to exploit the ontological relations of the model. OBIE systems seem to be capable of exploiting better the hierarchical ontological structures expressed rather than the event-based ontological models. 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 to identify 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 rule to annotate textual inputs with respect to ontological entities enabling semantic interoperability of annotations based on attributes that link textual representations to knowledge structures such as the EH Thesauri.

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 more complex JAPE rules have been made 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 managed to annotate 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 the grey literature contents and the semantic annotations that have been created during processing of the corpus collection. The objective of the third stage was to use the resultant XML files for making the semantic annotations available in simple form 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 is making use of the DOM XML for processing the XML files and revealing the annotations of documents while it is making use of a MySQL database server to store relevant thesauri structures. 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. Another early evaluation and visual inspection mechanism have been deployed in Andronikos web portal. A search engine indexing algorithm provided by the open source FDSE project has been deployed in the portal to index the web-pages of the semantic annotations and the full text version pages stored in the XML files. The search engine is then used to retrieve results from both indexes to visually inspect their ability to respond for the same search terms. 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 formulation 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 current restrictions 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 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. 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 [14].

VI.CONCLUSION

Today available semantic technologies promise to close the gap between formal knowledge structures and textual representations enabling new access methods to information [4][5]. 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 to enable 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 to reveal the semantic attributes of textual elements and to associate them to 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 to assist 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 required for further exploitation of the technique and adoption of formal evaluation methods for assessing the performance of the method in measurable terms.

REFERENCES

[1] Marchionini G, 1995 *Information Seeking in Electronic Environments* Cambridge: Cambridge University Press, NY, USA

[2] Smeaton A.F., (1997.). *An Overview of Information Retrieval* In: M. Agosti (ed),Information retrieval and hypertext, Kluwer Academic Publishers, p.3-20

[3] Jansen B, Spink A. (2006) *How are we searching the world wide web?* A comparison of nine search engine transaction logs. Information Processing and Management 42(1):pp. 248-263

[4] 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.

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

[6] 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)

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

[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] 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

[12] 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

[13] 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.

[14] 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