Semantic Annotation for Indexing Archaeological Context: A Prototype Development and Evaluation

Andreas Vlachidis[1], Douglas Tudhope[1]

[1] Hypermedia Research Unit, Faculty of Advanced Technology, University of Glamorgan, Pontypridd, CF37 1DL, Wales, UK
avlachid@glam.ac.uk, dstudhope@glam.ac.uk

Abstract. The paper discusses the process of developing Semantic Annotations, a form of metadata for assigning conceptual entities to textual instances, in this case archaeological grey literature. The use of Information Extraction (IE), a Natural Language Processing (NLP) technique is central to the annotation process. The paper explores the use of Ontology Oriented Information Extraction (OOIE) methods for the definition of rich semantic-aware indices of archaeology documents. The annotation process follows a rule-based information extraction approach using GATE. In particular the report discusses a prototype development that adopts the core ontology, CIDOC CRM, together with an English Heritage archaeological extension, to inform and direct the information extraction effort. The prototype evaluation, supports the assumptions made, about the capability of the method to construct rich indices of grey literature documents empowered by Semantic Annotations.

Keywords: Natural Language Processing, Ontology Based Information Extraction, Semantic Annotation, GATE, Digital Archaeology

1 Introduction

The complexity of human language results in a challenging environment for computations to provide solutions for the whole range of language related processes. On the other hand, advances in language engineering and computer technology of the past decades have made it possible for modern computer systems to perform natural language engineering tasks that previously were impossible to build and execute.

Information Extraction (IE) is a Natural Language Processing (NLP) technique which analyses a textual input and produces structured textual output of information capable for further manipulation. Such data manipulation can be directed towards automatic database population, machine translation, term indexing analysis and text summary generation. [1][2][3][4].

The fundamentally different role of IE does not compete with Information Retrieval (IR); on the contrary the potential combination of the two technologies promises the creation of new powerful tools in text processing. In particular, IR could benefit from the construction of sensitive indices of extracted information closer related to the "actual meaning" of a given text [1].

1.1 The Aims of the IE Prototype Development

The paper discusses the details of a pilot study, which prototyped both the development and evaluation methods of an Information Extraction system aimed at the delivery of semantic annotation metadata. The main objective of the pilot study was to explore and evaluate the capability of the CIDOC CRM [5] and CRM-EH [6] ontologies for modelling and resolving free text information from grey literature documents into semantic annotation. In order to accomplish the above objective, the pilot study investigated the capacity of the GATE [7] framework to accommodate the task of information extraction of grey literature documents of archaeological reports with respect to the above ontologies. Earlier results of the prototype development effort have confirmed the feasibility of GATE and JAPE grammars to support an initial IE task targeted at archaeological grey literature reports [8].

The prototype development explored the flexibility of GATE for modification and adaptation to a chosen semantic annotation task. The adaptation was concerned with the ability of JAPE grammar rules to target CIDOC CRM and CRM-EH concepts, as well as the capability of GATE gazetteers to accommodate resources like thesauri and glossaries which contain terms that enjoy unique terminological references. In addition, the study applied a pilot evaluation method for assessing the overall performance of prototype information extraction system. The evaluation method was informed by literature and followed established evaluation measurements for assessing the performance of semantic annotation systems. The delivered semantic annotations contribute to the STAR project which aims to achieve semantic interoperability over diverse archaeological resources [9]

1.2 The Platform of the IE Prototype Pipeline

The prototype pipeline was developed in the GATE (General Architecture for Text Engineering) [10] environment, utilising hand crafted JAPE rules and exploiting domain vocabulary that was made available as gazetteer listings.

GATE is described as an infrastructure for processing human language, which provides the architecture and the framework environment for developing and deploying natural language software components [7]. Offering a rich graphical user interface, it provides easy access to language, processing and visual resources that help scientists and developers to produce natural language processing applications.

JAPE (Java Annotation Pattern Engine) [11] is a finite state transducer, which uses regular expressions for handling pattern-matching rules. Such expressions are at the core of every rule-based IE system aimed at recognising textual snippets that conform to particular patterns while the rules enable a cascading mechanism of matching conditions that is usually referred as the IE pipeline. JAPE grammars are constituted from two parts; the LHS (Left Hand Side) which handles the regular expressions and the RHS (Right Hand Side) which manipulates the results of the matching conditions and defines the semantic annotation outcome.

1.3 The Role of Semantic Annotation

Semantic Annotation is the process of tying ontological definitions to natural text by providing class information to textual instances [12]. Described as a mediator platform between concepts and their worded representations, Semantic Annotation as metadata can automate the identification of concepts and their relationships in documents. It is proposed that a mechanism responsible for connecting natural language and formal conceptual structures could enable new information access methods and enhance existing ones.

Semantic annotation enriches documents, enabling access on the basis of a conceptual structure. This aids information retrieval from heterogeneous data enabling users to search across resources for entities and relations instead of words. Semantic Annotation has the potential to bridge the gap between natural language text and formal knowledge expressed in ontologies, as evident from a number of IE projects [13].

1.4 The CIDOC CRM – EH ontology

Ontologies are 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' [14][15]. The size and the scope, defines whether an ontology is called light-weight, core or upper level but all ontologies model a particular reality.

Ontological concepts can enrich information retrieval tasks by facilitating rich, semantic information seeking activities, both during query formulation and during retrieval. Inferences across diverse sources are supported by ontological structures, which are capable of mediating retrieval from heterogeneous data resources [16]. In addition, ontologies can be incorporated both in rule-based and machine-learning information extraction tools for supporting their semantic annotation operation. Usually such information extraction systems are described as ontology based (OBIE) or ontology oriented (OOIE), depending on the level of ontology engagement [17].

The CIDOC CRM is ISO standard ('ISO 21127:2006) core ontology for cultural heritage information aimed at enabling information exchange between heterogeneous resources by providing the required semantic definitions and clarifications. The CRM is the result of 10 years effort by the CIDOC Documentation Standards Working Group [18]. It is a comprehensive semantic framework designed to promote shared understanding of cultural heritage information.

The extended CRM model CRM-EH, is developed by the English Heritage EH, an organisation that has a major role in the dissemination of standards in cultural heritage domain both at national and international level. The extension resulted from the need to provide a common ground of shared meanings for what has been described as '*an archipelago of diverse, specialised and rather isolated and independent information systems and databases*' of the current archaeological systems [19]. The extended ontology comprises 125 extension sub-classes and 4 extension sub-properties. Based on the archaeological notion of context, modelled as place, the CRM-EH describes entities and relationships relating to a series of archaeological events, such as

stratigraphic relationships and phasing information, finds recording and environmental sampling.

1.5 Semantic Technologies for Archaeological Resources (STAR) project

The Semantic Technologies for Archaeological Resources (STAR) 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 supports the 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 [9].

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 data elements [20].

The project developed a CRM-EH based search demonstrator which cross searches over disparate datasets (Raunds Roman, Raunds Prehistoric, Museum of London, Silchester Roman and Stanwick sampling) and a subset of archaeological reports of the OASIS grey literature corpus. Also the project delivered a set of web services for accessing the SKOS[1] terminological references and relationships of the domain thesauri and glossaries which are employed by the project.

1.7 OASIS Grey Literature Reports

The term grey literature, it is used by librarians and research scholars to describe a range of documents and source materials that cannot be found through the conventional means of publication. Preprints, meeting reports, technical reports, working papers, white papers are just a few examples of grey literature documents which are not always published by conventional means. The expansion of the Web and the advent of sophisticated workstations increased the possibilities for disseminating information on a large scale. Thus the need for solutions targeted at accessing information with the volume of available grey literature documents is becoming more and more apparent [21].

A considerable volume of grey literature documents falls within the scope of the STAR project, constituting a valued resource 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

[1] Simple Knowledge Organization System (SKOS) is a standard language built upon RDF(S)/XML W3C technologies for the formal representation of knowledge organization systems, such as thesauri.

during and after major archaeological fieldwork. Integration of grey literature in STAR is intended for enabling cross-searching capabilities between datasets and grey literature documents, with respect to the semantics defined by the adopted CRM-EH ontology.

The collection of grey literature documents (corpus) that concerns the prototype development originates from the Online AccesS to the Index of archaeological investigations (OASIS) project [22]. The OASIS project is a joint effort of UK archaeology research groups, institutions, and organisations, coordinated by the Archaeology Data Service (ADS) [23], University of York, aiming to provide an online index to archaeological grey literature documents.

2 The Prototype Development of IE Pipelines

Two separate information extraction pipelines were developed to address particular objectives of the information extraction task. Both contribute to the main aim of the provision of semantic indexing with respect to the CRM-EH ontology.

The first pipeline (pre-process) is intended to reveal commonly occurring section titles of the grey literature documents and to extract the summary sections of grey literature documents. Summaries were identified from archaeology experts as important document sections containing rich information worth targeting by the main ontology oriented information extraction phase.

The use of the ontologies was examined and explored by the second, information extraction phase, which aimed at identifying pieces of information from grey literature documents, which could be associated with CRM and CRM-EH ontological entities. In particular, the pipeline explored the potential of the ontology to inform the construction process of the JAPE rules and its capacity to assign ontological definitions to the delivered semantic annotation metadata.

2.1 Pre-Processing Corpus Collection

The pre-processing phase (Figure 1) employed domain neutral information extraction techniques for the identification of specific document sections, which were then used as input at later stages of the prototype IE pipeline. During project discussions, archaeology experts suggested the exclusion of Headings and Table of Contents (TOC) from the semantic annotation process. It was made clear that such sections enjoy limited ontological commitment to CIDOC CRM, since they do not make use of terms in a rich discussion setting but instead use terms in isolation. Detecting headings also serves the purpose of revealing different document sections, such as summary sections, which contain rich discussion worth revealing. Heading annotations are used by the pipeline as input to detect the beginning and the end of each summary section.

The identification of the heading spans is based on a collection of eight different pattern-matching rules. Two rules annotate heading areas that commence with a numerical prefix followed by a capitalised or upper initial word, which might be

followed by more words not necessarily in capital or upper initial case, such as "*3.1 Prehistoric phase*". Another set of rules targets single worded headings that have upper initial or capitalised case and do not commence with numerical prefixes, such as "*Introduction*". Also a specific set of rules targets headings that are followed by a sequence of dots and a number, which are very frequently found in table of contents.

The identification of TOC is based on a simple pattern that joins four or more previously identified Heading annotations together. Similarly, identification of the summary sections is also based on a simple JAPE grammar, which matches sections wrapped between two heading annotations. The first heading annotation contains any of the words; "summary", "abstract" or "overview" independently of their case and the second heading annotation is simply the next occurrence of a heading in the document

2.1.1 JAPE Rules of the Pre-Processing Pipeline

As an example, consider the rule for matching headings beginning with a numeral. The rule matches phrases which commence with numbers like 1, 1., 1.1, 1.1.1. etc. followed by a non lowercase word Token, which is then followed from any number of Tokens including sequence of Dots (previously identified) until the end of line (EL) token.

```
({BL, Token.kind==number, Token.length <= 2}
(({Token.string == "."})?
({SpaceToken.kind == space})?
({Token.kind == number, Token.length <= 2})?)*
({SpaceToken.kind == space})+
({Token.orth !="lowercase", Token.kind == word})
({Token.kind==word}|{Token.kind==number}|
{Token.kind==punctuation}|{SpaceToken.kind==space}|
{Dots})*{EL})
```

The following rule matches document summary sections which commence with a heading annotation of the type *Summary* (matched by a previous rule) and ends with the next available heading annotation of the document. It is possible to identify large chunks of text by configuring the rule to process only the heading annotations not Tokens or other annotation types, simplifying this way the grammar of the pattern.

```
{Heading.type=="Summary"}
{Heading}
```

2.2 Domain Oriented Information Extraction Pipeline

The domain-oriented pipeline (Figure 1) extracts specific archaeological information utilising the available English Heritage terminology resources and the domain ontologies, CIDOC CRM and CRM-EH. The choice of ontological entities targeted by the process is based on project discussions with English Heritage and specifically with project archaeological collaborator, Keith May, and the study of available use

case scenarios. It was decided that the prototype system should focus on the extraction of the following concepts:

a) E19 Physical Object described as "*items having physical boundaries that separate them completely in an objective way from other objects*"
b) E49 Time Appellation described as "*appellation of all forms of names or codes, such as historical periods, and dates, which are characteristically used to refer to a specific temporal extend that has a beginning an end and a duration*"
c) E53.Place with emphasis on EHE0007.Context described as "*Spatial elements that constitute an individual archaeological unit of excavation including both primitive contexts and larger groupings of contexts*"

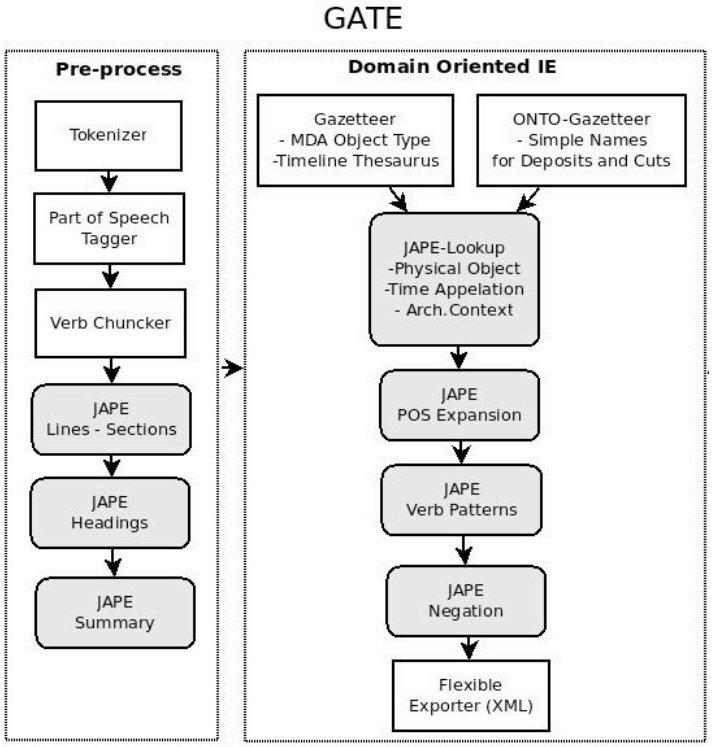

Fig. 1. The Information Extraction pipeline developed in GATE. Two separate exercises are shown here the Pre-processing and the Domain Oriented IE. Bespoke JAPE rules shown in Grey boxes, white boxes are used for GATE modules

2.2.1 Terminology Resources for Information Extraction

A range of different terminology resources such as thesauri and glossaries were made available to the STAR project by English Heritage, which were transformed to SKOS XML files (Skosified) in a previous exercise [24]. In addition, the "Skosified"

resources were transformed into GATE gazetteer listings using XSLT transformation templates. The pilot system utilizes three particular resources; the Archaeological Object Type thesaurus, the Time-line thesaurus and the glossary Simple Names for Deposits and Cuts [25]. In addition, the pipeline exploits an RDF version of the CRM-EH ontology which is imported into GATE using the OWLIM ontology repository plug-in of GATE.

All gazetteer resources were assigned a *Major Type* and a *Minor Type* property, which were accessed by JAPE grammars for the definition of Lookup rules. The *Major Type* property was used for declaring the conceptual definition of each resource. Hence, all terms originating from the Archaeological Object Type thesaurus were given a *Major Type* "Physical Object", the terms of the Time-line thesaurus were given a *Major Type* "Time Appellation" while the glossary populated the OWLIM repository with instances of the CRM-EH class EHE0007 Context. Each individual term was also assigned a *Minor Type* property corresponding to its *skos:concept* terminological reference, maintaining in this way a link between GATE gazetteer and SKOS structure.

2.2.2 JAPE Rules of the Domain Oriented Information Extraction Pipeline

The prototype pipeline implemented fifteen different JAPE grammars for identifying the three main ontological concepts (Physical Object, Archaeological Context, and Time Appellation). The grammars exploited the *Major Type* gazetteer property for assigning the corresponding ontological reference to the matches, with the exception of the matches of Archaeological Context which instead of Major Type property used the CRM-EH class property. This is because the Lookup mechanism for Archaeological Context experimented with the use of the OWLIM ontology plug-in.

Additional rules were used for extending the initial Lookup annotations to include meaningful moderators. In the case of Time Appellation, two gazetteer listings were used for expanding over prefix terms (Earlier, Later, etc.) and suffix terms (Period, Century, etc.)

The following rule matches three different cases of Time Appellation expansion. a) Expansion towards prefix and suffix i.e. "Early Roman Period", b) Expansion only towards prefix i.e. "Early Mediaeval" and c) Expansion only towards suffix i.e. "Prehistoric period"

```
({Lookup.minorType==Date_Prefix}{TimeAppellation}
{Lookup.minorType==Date_Post})|
({Lookup.minorType==Date_Prefix}{TimeAppellation}|
{TimeAppellation}{Lookup.minorType == Date_Post})
```

Moreover, JAPE patterns were employed to identify pairs of Lookups, such as Time Appellation and Physical Object i.e. "Roman Pottery" or Time Appellation and Archaeological Context, i.e. "Mediaeval Deposit". This last approach was elaborated further by the definition of JAPE patterns that matched linguistic evidence of combinations between Lookups and verb phrases in the form of <Lookup><verb><Lookup>. Such patterns were aimed at matching combinations between Time Appellation and Physical Object as for example "...coins dating to

Roman period...", Time Appellation and Archaeological Context as for example "...pits are of prehistoric date...", and Archaeological Context and Physical Object as for example "…pits containing pottery...". This above pattern-matching approach was aimed at supporting the assumption that text phrases carry information which describes relations between CRM-EH entities and that linguistic evidence in form of pattern matching rules can be employed to extract such textual instances.

The following rule matches phrases that connect Lookup annotation via verb phrases i.e. "pits containing pottery"

```
{Context}({Token.kind==word}|{Token.category==","})*
{VG}({Token.kind==word}|{Token.category==","})*
{PhysicalObject}
```

3. Evaluation of Semantic Annotations

The effectiveness of Information Extraction systems is measured in *Recall* and *Precision* rates. The measurement units originate from the IR domain but they have been redefined during the Machine Understanding Conference MUC to reflect matching and mismatching of the information extraction process [26].

The evaluation task aimed at measuring the performance of the prototype information extraction mechanism with regards to the concepts of Time Appellation, Physical Object and Archaeological Context. The task had a largely formative and less summative character, aiming not just to evaluate the performance of the prototype system but also to suggest the necessary development improvements that have to be taken on board by a full scale Information Extraction exercise.

For the purposes of the evaluation, a manually annotated versions of the intended IE results was created and made available to the GATE Corpus Benchmarking Utility. Four versions of individual manual annotation sets were produced by four project members including the developer and an archaeology expert.

Since the major aim of the pilot evaluation was to inform a later larger scale IE the evaluation exercise did not conclude in a single definitive "gold standard" [27] version, which was not considered necessary for the purposes of the pilot study. Instead, the evaluation task used all four manual annotation versions in order to get a pluralistic view for the system performance, informed by the differences of manual annotation between individual sets.

The summary sections which participated in the evaluation task were extracted during the pre-processing phase. The selected extracts originated from five archaeological Evaluation reports and five archaeological Excavation reports, which were identified by the archaeology expert as sections carrying rich and relevant information to the aims of the prototype evaluation.

3.2 Evaluation Results

A closer examination of the overall system's performance (Table 1) revealed encouraging results regarding Precision, Recall and their weighted average F-

measure. When including AV's version of manual annotation to the overall score the system's performance score improved even further. This was to be expected since AV was involved in the development of the system and so was more aware of the capabilities of the extraction mechanism regarding coverage of gazetteer resources and pattern matching rules. On the other hand, the system delivered some positive results against KM, who is an archaeology expert involved in the definition of the CRM-EH ontology and so his judgment is considered to be more definite and closer to the ontological definition than the rest of annotators.

Examining the performance of the system against manual annotators reveals that there is some basic agreement between annotators about the system performance. Excluding AV, the system delivered an average fMeasure score of 56%, marking the system's ability to target specific concepts with some success.

Table 1 System's performance against the gold standard Annotations

	AV	**CB**	**DT**	**KM**
Precision	0.85	0.68	0.73	0.69
Recall	0.85	0.69	0.61	0.71
F-Measure :	0.76	0.56	0.56	0.56

Table 2 System's performance for three ontological entities showing differentiation between annotators

	Entity	**Correct**	**Missing**	**Precision**	**Recall**	**F-Measure**
AV	E49	51	5	0.99	0.90	0.94
	E19	14	4	0.53	0.75	0.62
	EHE007	56	2	1.00	0.96	0.98
CB	E49	46	20	0.93	0.68	0.79
	E19	13	20	0.50	0.40	0.44
	EHE007	35	66	0.62	0.35	0.44
DT	E49	44	16	0.90	0.71	0.80
	E19	9	19	0.41	0.35	0.37
	EHE007	35	78	0.66	0.31	0.42
KM	E49	45	35	0.93	0.55	0.70
	E19	10	17	0.42	0.39	0.40
	EHE007	31	96	0.60	0.25	0.36

The system's performance on individual ontological entities reflects the differentiation between individual annotation sets. The system performed well against E49 Time Appellation entities delivering high *Precision* varying from 90% to 99% and *Recall* rates varying from 55% to 90% (Table 2). Precision was also good for EHE0007 Archaeological Context entities (60% to 66%). However, Recall for the

same entity type was low, varying from 25% to 35% due to the limited coverage of the ontology list of instances for this particular entity. The coverage of the gazetteer listing for the E19 Physical Object entity was also problematic affecting the overall performance of the system for this particular type. The system for Physical Object annotations, delivered a precision score around 50%, indicative of the volume of false positive matches identified.

4 Conclusions

The prototype development has reached its aim of implementing a prototype IE system, capable of extracting concepts with respect to a given domain ontology and generating rich semantic annotations of grey literature documents. The initial evaluation results are encouraging and have revealed the capacity of the method for identifying rich textual instances that correlate to a set of ontological entities and properties. Extraction of ontological phrases that combine more than one ontological entity also looks promising. Such phrases carry the potential of extracting CRM-based Event type entities. In addition, the results of the pilot study suggest that utilization of the hierarchical relationships of the available thesauri is required by a full-scale system. A sophisticated exploitation of thesaurus relationships could benefit the IE outcome by enabling a selective use of the terminology resources that does not harm recall by using too little or precision by using too much of the available vocabulary. Further elaboration of the method is required for revealing further the capabilities of NLP techniques to provide rich semantic indices at an operational level.

Acknowledgements. The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). Thanks are due to Phil Carlisle & Keith May (English Heritage), Ceri Binding (University of Glamorgan), Renato Souza (Universidade Federal de Minas Gerais, Brazil).

References

1. Cowie J, Lehnert W.: Information extraction. Communications ACM, vol 39(1), pp.80--91. ACM, New York (1996)
2. Lewis D, Jones K.: Natural language processing for information retrieval. Commun. ACM, vol. 39(1), pp.92--101. ACM, New York (1996)
3. Moens M.F: Information Extraction Algorithms and Prospects in a Retrieval Context. Springer, New York (2006)
4. Gaizauskas R, Wilks Y.: Information extraction: beyond document retrieval. Journal of Documentation, vol. 54(1), pp.70--105. Emerald, Bradford (1998)
5. CIDOC-CRM, http://www.cidoc-crm.org/
6. CRM-EH, http://hypermedia.research.glam.ac.uk/resources/crm/
7. Cunningham H, Maynard D, Bontcheva K, Tablan V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics ACL'02 (2002)

8. Vlachidis A, Binding C, May K, Tudhope D.: Excavating Grey Literature: a case study on the rich indexing of archaeological documents via Natural Language Processing techniques and Knowledge Based resources. ASLIB Proceedings journal, vol. 62 (4&5), pp.466--475, (2010)
9. Tudhope D, Binding C, May K.: 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, associated with 5th European Semantic Web Conference, pp. 88--99. Tenerife (2008)
10. General Architecture for Text Engineering GATE, http://gate.ac.uk/
11. Cunningham H, Maynard D, Tablan V.: JAPE a Java Annotation Patterns Engine (Second Edition). Technical report CS--00--10, University of Sheffield, Department of Computer Science (2000).
12. Bontcheva K, Cunningham H, Kiryakov A, Tablan V.: Semantic Annotation and Human Language Technology. Semantic Web Technology: Trends and Research in Ontology Based Systems. John Wiley and Sons, Sussex (2006)
13. Uren V, Cimiano P, Iria J, Handschuh S, Vargas-Vera M, Motta E, Ciravegna F.: 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, vol. 4(1), pp 14--28. Elsevier, Amsterdam (2006)
14. Guarino, N.: Formal Ontology and Information Systems, In N. Guarino, (ed.), Formal Ontology in Information Systems, IOS Press, pp.3–15 (1998).
15. Wilks Y.: The Semantic Web as the apotheosis of annotation, but what are its semantics? Intelligent Systems, vol. 23(3), pp.41–49. IEEE Press, New York (2008)
16. Kiryakov A, Popov B, Terziev I, Manov D, Ognyanoff D.: Semantic annotation, indexing, and retrieval. Web Semantics: Science, Services and Agents on the World Wide Web, vol. 2(1), pp. 49–79. Elsevier, Amsterdam (2004)
17. Bontcheva K, Duke T, Glover N, Kings I.: Semantic Information Access. In Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems John Wiley and Sons. Sussex, (2006)
18. Crofts N, Doerr M, Gill T, Stead S, Stiff M.: Definition of the CIDOC Conceptual Reference Model. http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Mar09.pdf
19. Cripps P, Greenhalgh A, Fellows D, May K, Robinson D. E.: Ontological Modelling of the work of the Centre for Archaeology. CRM – EH model diagram (2004) http://cidoc.ics.forth.gr/docs/AppendixA_DiagramV9.pdf
20. STAR project, http://hypermedia.research.glam.ac.uk/kos/star/
21. Debachere M.C.: Problems in Obtaining Grey Literature. IFLA Journal, vol. 21(2) p 94. IFLA, Edinburgh (1995)
22. Online AccesS to the Index of archaeological investigations OASIS, http://oasis.ac.uk/
23. Archaeology Data Service ADS, http://archaeologydataservice.ac.uk
24. Binding C, Tudhope D, May K.: 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 Digital Libraries, Aarhus (2008)
25. EH National Monuments Records Thesauri http://thesaurus.english-heritage.org.uk/
26. Grishman R, Sundheim B.: Message Understanding Conference-6; a brief history. Association for Computational Linguistics, pp.466–471. New Jersey (1996)
27. Maynard D, Peters W, Li Y.: Metrics for Evaluation of Ontology-based Information Extraction. In Proceding of WWW 2006 Workshop on "Evaluation of Ontologies for the Web", (2006).