Automatic Metadata Generation in an Archaeological Digital Library: Semantic Annotation of Grey Literature

Andreas Vlachidis
Hypermedia Research Unit,
University of Glamorgan,
CF371DL Pontypridd, UK
Email:
avlachid@glam.ac.uk

Ceri Binding, Douglas Tudhope
Hypermedia Research Unit,
University of Glamorgan,
CF371DL Pontypridd, UK
Email: {cbinding,
dstudhope}@glam.ac.uk}

Keith May
English Heritage,
Fort Cumberland,
Fort Cumberland Road,
Southsea PO4 9LD
Email: keith.may@english-heritage.org.uk

***Abstract*—This paper discusses the automatic generation of rich metadata for semantic search of reports of archaeological excavations. An extension of the CIDOC CRM for the archaeological domain acts as a core ontology. This enables cross search between diverse excavation datasets and 'grey literature' excavation reports originating from the Archaeological Data Service OASIS library. Rich metadata is automatically extracted from the reports, directed by the CRM, via a three phase process of semantic enrichment employing the GATE toolkit. This is expressed as XML annotations coupled with the reports and also as RDF metadata, both represented as CRM entities, qualified by SKOS archaeological concepts. A web portal delivers the annotated XML files for visual inspection while the STAR research demonstrator offers unified search of excavation data and grey literature in terms of the conceptual structure. Initial evaluation results show operational precision and recall rates for three different semantic expansion configurations of the system.**

I. Introduction

THE recent growth in semantic technologies has made possible Digital Library developments that take advantage of the potential of rich semantic contextualisation, based on conceptual models. Europeana is a prominent example, a digital library linking more than 6 million digital items from the cultural and heritage domain [1]. It aims to deliver a significant semantic enrichment to its linked digital objects, via the Europeana Data Model (EDM) which subsumes the CIDOC Conceptual Reference Model (CRM) amongst other metadata models [2].

The CRM is an international standard (ISO21127:2006) semantic framework, aiming to promote shared understanding of cultural heritage information [3, 4]. Extensibility is an important aspect; a finer granularity of detail can be expressed for domain purposes while still retaining interoperability at the core CRM level. A particular extension of CRM in the archaeology domain is the English Heritage (EH) CRM-EH ontology, which comprises 125 extension sub-classes [5].

As part of the general move to making data publicly available for reuse, archaeology has seen the emergence of digital libraries that feature excavation datasets along with textual reports, for example the Archaeology Data Service (ADS). In the UK, the majority of excavation reports come in the form of grey literature. The term "grey literature" refers to publications that are not formally published such as on-line documents, excavation reports, watching briefs etc. The OASIS (Online AccesS to the Index of archaeological investigations) grey literature library is a joint effort of UK archaeology organisations, coordinated by the ADS [6], aiming to improve the communication of fieldwork results to the wider archaeological community. However, the digital libraries of grey literature reports and excavation datasets are not meaningfully connected.

A. Aims and Overviews

This paper discusses a novel contribution in the automatic generation of rich metadata, driven by the CRM-EH. The STAR (Semantic Technologies for Archaeological Resources) project employs the CRM-EH as a core ontology to provide a conceptual framework for semantic contextualisation [7, 8]. In collaboration with EH, STAR has developed methods for linking digital archive databases, vocabularies and OASIS excavation reports for the purposes of semantic cross search. A broad overview of the STAR Demonstrator and its cross searching capabilities is available from a previous paper which, however, does not discuss the NLP methods in any detail.

This paper focuses on the methods developed for automatically extracting rich metadata from the grey literature, directed by the CRM-EH, via a process of semantic enrichment employing the General Architecture for Text Engineering (GATE) toolkit [9]. Two web applications use the resultant semantic annotations. The Andronikos web portal delivers the annotated XML files as hypertext documents for visual inspection of the information extraction results. The STAR research demonstrator offers a unified searching of both data and grey literature in terms of the core ontology.

II. Semantic Annotation

There has been a considerable amount of effort dedicated over the last years in the design and development of Semantic Annotation Platforms [10, 11, 12]. A detailed description of all available tools and platform expands beyond the scope of this paper.

KIM is a well known application which uses metadata in form of semantic annotation to support information retrieval tasks [13]. The application uses KIMO, an upper-level onto-

logy consisted of 200,000 instances of various generic type entities (place, location, people, etc.). In order to satisfy semantic queries the platform employs the SESAME repository of RDF triples and a modified version of the Lucene search engine.

Archaeotools is an archaeology oriented application aimed at creating semantic infrastructure for archaeology research [14]. The project generated a faceted classification for the metadata records of the UK archaeological sites and monuments and the associated antiquarian and grey literature reports, maintained by the ADS. The four facets of the classification are What, Where, When and Media. This builds on previous experience with faceted browsing in the ADS Archaeobrowser service.

While the Archaeotools project does open conceptual access to grey literature documents, it employs a broad classification system which cannot exploit the potential of semantic annotation metadata for answering detailed ontology driven semantic queries, which may combine ontological entities in rich indexing phrases. Concerning excavation data, Archaeotools is based on the metadata of the underlying reports, rather than opening access to the datasets themselves, which is STAR's focus. Essentially, STAR complements the Archaeotools approach by offering more specific semantic access, both to data and to the grey literature.

Similarly KIM and related semantic annotation applications are based on upper-level ontologies, which cannot address detailed, domain oriented information needs. For the purposes of the STAR project, the CRM-EH extension of the CRM was adopted as a domain oriented ontology, in order to satisfy comprehensive semantic access to archaeological data and reports. Use of the CRM for rich semantic annotations of text documents has been explored via intellectual process. This has the potential for producing very fine grained annotation of specific, important documents, for example as part of detailed Text Encoding Initiative mark-up [15]. Inevitably, this process will be resource intensive over a large corpus. This paper reports on an investigation of automatic methods for generating rich metadata that connects concepts via CRM events and properties.

III. The Process of Semantic Enrichment

The discussion on the process of Semantic Enrichment is divided into the three broad phases, each one subdivided into various sub-tasks (pipelines). The initial phase pre-processes the grey literature and vocabulary resources, while the second phase identifies domain concepts in context. The final phase transforms GATE annotations to semantically enriched documents in form of XML annotations coupled with the grey literature reports and decoupled RDF representations of metadata. Both outputs are expressed as CRM entities, qualified by SKOS[1] archaeological vocabulary concepts [16]. This process builds on a previous pilot study, which explored the semantic indexing of grey literature documents using the CRM [17]. The current pipeline significantly advances the basic (look-up based) information extraction in the pilot via information extraction methods that exploit the semantic relationships of contributing knowledge resources, invoking sophisticated rules for dealing with negation detection and word sense disambiguation, and targeting the specificity of the CRM-EH with pattern matching rules informed by corpus analysis bottom up strategies. Another contribution of the work reported here is the automatic generation from the textual reports of rich semantic metadata, combining different ontology entities, expressed as RDF for purposes of cross search with the STAR datasets.

A. Underlying Architecture and Knowledge Organization System Resources

A popular open source Language Engineering platform that can accommodate the task of IE is the General Architecture of Text Engineering (GATE). GATE is described as a development environment for developing and deploying natural language software components [9]. The architecture integrates the Java Annotation Pattern Engine (JAPE), enabling the construction of regular expressions in the form of JAPE rules. The architecture makes available a range of language processing resources, such as the Tokenizer, Sentence Splitter and Part-of-Speech tagger, as part of the default application ANNIE (A Newly New Information Extraction System).

B. Underlying Architecture and Knowledge Organization System Resources

Gazetteers are sets of lists, sometimes containing the names of entities such as cities, day of the week, etc. In GATE, gazetteer listings are used to find occurrences of terms in free text, often supporting named entity recognition tasks. GATE gazetteers are not necessarily flat, which means that enlisted terms can enjoy attributes which in turn can be invoked by JAPE rules for the construction of sophisticated patterns. Four thesauri and five glossary resources are incorporated as GATE gazetteers in the process of semantic enrichment for identifying occurrences of various conceptual entities. The thesauri are the MDA Object Types, the Monument Types, the Main Building Materials and the Time-line Thesaurus [18]. The glossary resources used in the process are the Simple Names for Deposits and Cuts, the Find Type Index, the Material Index, the Small Finds and the Bulk Find Material glossary. All resources were expressed in SKOS format for the purposes of the STAR project [19].

The glossary resources contain a small set of concepts, which are highly relevant to archaeological excavations. On the other hand, the thesauri resources contain a large set of cultural heritage concepts, being developed for more general purposes. Exploiting the whole range of thesauri resources would expand the process of enrichment to concepts that are not very relevant to excavation reports or datasets. Therefore, an optimum range of thesauri concepts is needed. This paper reports on the novel use of overlapping concepts between glossary and thesauri as entry points to the thesaurus structures. Closely related concepts in the thesauri can also be relevant. Thesaurus semantic relationships are exploited for expanding from the entry point across thesauri

[1] Simple Knowledge Organisation Systems (SKOS) is a standard set of languages built upon standard RDF(S)/XML W3C technologies for formal representation of structured controlled vocabulary systems such as thesauri, classification schemes

areas relevant to the task of semantic enrichment. The "skosification" of GATE gazetteers (i.e. transformation of SKOS thesauri to gazetteers allowing use of thesauri properties as gazetteer attributes) is a novel contribution of this paper. In order to enable this semantic expansion within GATE, JAPE rules were written to exploit narrower and broader thesaurus relationships.

C. Pre-processing Phase

There are two main pre-processing components: preparing knowledge resources for use within GATE and identifying the basic section structure of each OASIS document.

a) Transforming Thesauri to GATE gazetteers

In order to work with existing GATE structure, SKOS thesauri and glossary resources had to be transformed to GATE gazetteers. This was achieved via the use of XSLT templates. The templates exploit SKOS properties for adding attributes to gazetteer terms, which can be used by JAPE rules. It is important that the rules are capable of traversing through a thesaurus hierarchy, in order to produce matches that achieve a semantic expansion. Since JAPE is essentially a pattern matching rule engine, parametrisation of gazetteers is required to enable the semantic expansion of SKOS concepts (with their unique identifiers) within GATE.

Consider the following case; *Container (by function) >Food and Drink Serving Container > Drink Serving Container > Jug > Knight Jug*. A JAPE rule capable of exploiting a narrower term relationship is able to semantically expand on all immediately narrower concepts. Therefore, a gazetteer attribute was built during the transformation from SKOS to GATE gazetteer to reflect the path of unique SKOS identifiers from the concept to the top of its hierarchy. For example:

```
KnightJug@skosConcept=
149773@path=/101601/101204/101340/101023
```

A simple JAPE rule can then exploit the above gazetteer attributes by matching all terms that contain a skosConcept and a path attribute of a particular reference; for example the SKOS reference 101023 matches all concepts in the gazetteer within the *Container* hierarchy. The XSLT transformation also takes into account SKOS alternative concept labels (thesaurus non-preferred terms) and makes them available as gazetteer entries that have the same skosConcept and path attributes as their preferred label counterparts.

In addition during transformation, particular glossary concepts are given an extra attribute (skos:exactMatch) to accommodate the previously defined mapping between a glossary and a thesaurus. For example the concept *Hearth* of the glossary Simple Names for Deposits and Cuts is mapped to the concept *Hearth* of the thesaurus Monument Types class Archaeological Feature. This allows the potential for JAPE rules to optionally expand the concept *Hearth* to associated concepts within the Monuments thesaurus.

b) Identifying document structure

The pre-processing phase also uses domain neutral JAPE rules for identifying particular document sections for differential levels of priority in the subsequent phase of semantic enrichment.

The pre-processing phase was able to identify summary sections of the OASIS corpus. Summaries report on the main findings of an excavation report and thus semantic metadata deriving from summaries has particular relevance. On the other hand, sections such as headings and table of contents are currently considered less relevant in the semantic enrichment process because any references to domain entities are abstracted from their context. Tabular data are also currently excluded since they would require specialized treatment, being at a lower granularity of detail.

Heuristic rules are used to define JAPE patterns for the identification of document areas. Briefly, these patterns make use of syntactical evidence such as length of sentences, numerical commencement and use of letter case in order to identify the various sections. The pre-processing phase uses ANNIE modules to tokenize documents on a word level and to identify noun and verb phrases. Noun and verb phrases are used during the main IE phase for validating lookup generation, for example distinguishing the verb from the noun sense of the word *building*.

D. Main Knowledge-Based Information Extraction Phase

The second phase of semantic enrichment is dedicated to the main IE process aimed at the annotation of grey literature documents with conceptual and terminological references. A dedicated pipeline was developed within GATE - OPTIMA (Object, Place, TIme and MAterial). These four concepts were considered key metadata elements for the project's concern with archaeological excavations. They are the focus of the IE process, which uses a large number of JAPE patterns and utilises the gazetteer resources created by the previous pre-processing phase. The section highlights the main functionality of the pipeline but does not elaborate on the details, which fall out of the scope of this paper.

The first stage of the main IE pipeline invokes gazetteer resources and generates the initial lookup annotations. There are cases where a term can be found within two different knowledge resources, thus potentially having two different conceptual references. For example in the particular domain practice of archaeology, the term *brick* can refer to a material or to a physical object, as can terms such as, glass, stone, iron, gold etc. Therefore, the first stage of the pipeline generates two types of lookup annotations, single sense and multiple sense. Multiple sense lookup annotations are disambiguated at later stages.

The second stage of the pipeline validates the Lookup annotations and aligns annotations to CRM entities. Annotations that are not part of noun phrases and annotations, that are part of headings, table of contents and single worded phrases are suppressed. It is important to validate Lookup annotations, especially those that are not part of noun phrases, because gazetteer matches are invoked via a morphological analyser and matches are created on the root of words. This technique allows matches within a broader orthographical context, including singular and plural forms of matches, but also generates matches for verb senses having the same word root with noun senses that have to be sup-

pressed by the validation stage since Named Entity Recognition is targeted at nouns not at verbs. In addition, during this stage the pipeline performs negation detection over lookup annotations. The negation detection technique is based on NegEx algorithm which originates from the medical domain [20]. As part of the work, the NegEx algorithm was modified for application in the domain of archaeology. The next stage of the pipeline disambiguates multiple sense lookup annotations. The disambiguation technique is based on JAPE patterns that examine word pairs and Part-of-Speech input. Lookup annotations that cannot be disambiguated maintain all possible senses.

The last stage of the pipeline provides conceptual references to annotations with respect to the CRM-EH model. In order to accomplish annotation at the CRM-EH semantic level, the pipeline invokes a set of JAPE patterns that find rich phrases connecting the previously identified lookup entities. The construction of JAPE patterns is informed by the CRM-EH ontological definitions and the results of a bottom up corpus analysis, which identified the commonly occurring patterns that connect the four different types of entities.

The pipeline can be configured to perform in three different modes of semantic expansion; Synonym, Hyponym, and Hyperym expansion. Synonym expansion utilises the glossary resources and expands on the synonyms of glossary terms available in the thesauri resources. Hyponym is similar to the Synonym expansion in utilising the glossary terms and their synonyms but also traverses over the hierarchy of the thesaurus structures to include (transitively) all narrower available for the glossary concepts. The Hypernym expansion mode includes the above two modes of expansion and also exploits the broader concept relationships within the thesauri structures. In terms of volume, Hypernym expansion expands the semantic enrichment task to include the largest set of concepts, while Synonym expansion includes the smallest and most precise set of concepts. Clearly, there are recall/precision trade-offs associated with the different expansion modes which are examined by the evaluation.

E. Transformation of Semantic Annotations to RDF triples

The last phase of the semantic enrichment is the transformation of the semantic annotations produced in the previous phase to RDF triples. For this purpose, CRM-EH semantic annotations are exported initially from the GATE environment as XML documents. The GATE exporter produces annotations in the form of XML tags, which are coupled with the associated content. Each grey literature document has a unique name that constitutes a unique identification of the file within the OASIS corpus. This unique file name is used in conjunction with the unique *gateID* property of each annotation to create a corpus wide unique identifier for each individual annotation. In addition, the SKOS concepts assigned to the annotations are associated with underlying CRM Types, using the same project specific relationship (is_represented_by[2]) modelling the association asserted for data items (mapped to CRM) and SKOS concepts [19]. For example a semantic annotation of CRM-EH Context (a subclass of CRM Place) can be associated with the SKOS type, *pit*. This supports cross search between data and grey literature in terms of CRM and SKOS.

The transformation from XML files to RDF triples is based on the XML Document Object Model, using the scripting language PHP for building the transformation templates. The final RDF documents are decoupled from the content. The following example presents the details of an RDF transformation. Consider the following rich phrase "*pits were uniformly filled with large quantities of pottery*". The phrase can be modelled by a CRM-EH ContextFindDepositionEvent, connecting a find (*pottery*) with a context (*pit*). The RDF transformation would have the following structure:

```
  <crmeh:EHE1004.
ContextFindDepositionEvent
rdf:about="http://base#suff1-
6115.281105">
  <dc:source rdf:resource=
"http://base#suffolkc1-6115" />
  <dc:source rdf:resource=
"http://base#ehe0001.oasis" />
  <crm:P2F.has_type rdf:resource
="http://base#suff1-6115.281156" />
   <crm:P3F.has_note>
    <crm:E62.String>
     <rdf:value>
pits were uniformly filled with large
quantities of pottery
     </rdf:value>
    </crm:E62.String>
   </crm:P3F.has_note>
  <crm:P26F.moved_to rdf:resource
="http://base#suff1-6115.281155" />
  <crm:P25F.moved rdf:resource
="http:// base#suff1-6115.281158" />
</crmeh:EHE1004.
ContextFindDepositionEvent>
```

IV. Example Uses of Rich Metadata

The automatically produced metadata are utilised by two web applications, the STAR research demonstrator and the Andronikos portal [21, 22]. The STAR demonstrator uses the decoupled RDF files to support cross searching between grey literature documents and disparate datasets [23], in terms of the core CRM-EH conceptual model. A SPARQL engine supports the semantic search capabilities of the demonstrator, while an interactive interface hides the underlying model complexity and offers search (and browsing) for Samples, Finds, Contexts or interpretive Groups with their properties and relationships. On the other hand, the Andronikos portal uses the coupled XML files for constructing and delivering the semantic annotations in an easy to follow human readable format. While the portal was developed for project purposes to assist visual inspection of the information extraction outcome, it is seen as indicative of potential

[2] In the absence of a compelling alternative, the project specific relationship was adopted to facilitate subsequent transition to any emerging standard.

digital library applications where access to the semantically enriched text is desired.

The STAR Demonstrator makes use of the rich metadata for some forms of semantic search, building on CRM and SKOS unique identifiers. For example, searches are possible of the form: *Context of type X containing Find of type Y*. The screen-dump in Figure 1 shows a Context Find of type *"Animal Remains" within a Context of type "pit"*. The cross-search capability of the STAR demonstrator retrieves results from both datasets and grey literature reports.

For example, the search retrieves from Grey Literature an animal bone within a context of type pit; the original text was "the test pit produced a range of artefactual material which included animal bone (medium/large ungulate)". The semantic enrichment makes it possible for the STAR demonstrator to overcome lexical boundaries and to retrieve synonymous terms, as evident in the example of "A*nimal Remains*" where the term "*Animal bone*" is retrieved.

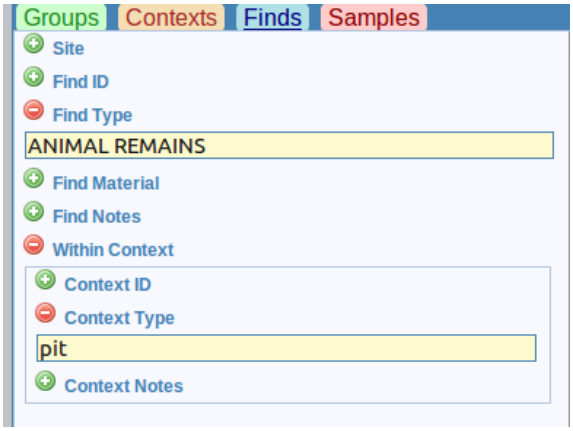

Fig. 1 The STAR demonstrator search of semantic metadata

The Andronikos web portal uses XML outputs of rich metadata for generating and linking HTML pages, which accommodate semantic annotations of grey literature documents. The portal makes it is possible to optionally expose particular document abstractions according to different application strategies. Thus, in certain cases, the Summary sections might be targeted (or prioritised) for retrieval as being strongly representative of a grey literature report. Alternatively, the most frequently appearing CRM entities (see Figure 2) in a report might be considered a useful entry strategy. Yet again, a cross search might be interested in any occurrences within grey literature reports of highly specific, rich CRM-EH annotations.

The CRM overview (Figure 2) show the most frequently used SKOS concepts for the CRM entities Time Appellation and Physical Object, in this case Roman and pottery. In addition, CRM-EH rich metadata such as *"3rd century pottery was recovered from its medium greyish brown silty sand fill"* reveal evidence that make connections between some frequently used SKOS metadata, in terms of the CRM-EH Deposition and Production events (subclasses of CRM events) relating to archaeological finds. In this case, a Context Find (*pottery*) is connected with a Time Appellation (*3rd century*) via the CRM-EH entity, Context Find Production Event. The Context Find (*pottery*) is also connected in the same phrase with an archaeological Context (*sand fill*), via the CRM-EH entity Context Find Deposition event[3]. CRM-EH Context Find is a subclass of CRM E19:Physical Object and Context is a subclass of CRM E53:Place. Using the above CRM-EH entities, a semantic application can make further inferences about the CRM entities; as for example to possibly connect the archaeological Context (sand fill) with the Time Appellation (3rd century), depending on the dating of any other finds within the same context. Generally in OASIS reports (ie. reports following analysis of all the finds) when a date is mentioned for a find, there is an assumption that the find's date has been taken as diagnostic of the context in which it was found. The rich metadata opens the possibility for very precise semantic queries based upon the connection of entities via the CRM events.

TERM	SKOS	Count
fired clay	#ehg027.5	40
plate	#96797	53
artefacts	#ehg020.7	55
tile	#ehg027.3	57
pottery	#ehg027.2	81

TERM	SKOS	Count
anglian	#136306	14
medieval	#134745	19
20th century	#134841	19
prehistoric	#134718	46
roman	#134738	216

Fig. 2 Frequent CRM entities (Time Appellation - Physical Object)

V. Evaluation

Performance of the pipeline was evaluated against *Recall* and *Precision* following an expert manual annotation evaluation. The evaluation task aimed to benchmark the performance of the information extraction mechanism for the concepts Physical Object, Place, Material, Time Appellation and their CRM-EH specialisations Context, Context Find and Context Find Material specialised by the CRM-EH events Context Event, Context Production Event and Context Deposition Event. A set of guidelines was provided to three archaeology experts for identification of phrases carrying rich meaning with regards to the targeted concepts. The resulting manual annotation sets, which are discussed in this paper, are an initial evaluation exercise, informing an ongoing larger scale 'gold standard' evaluation.

Calculation of the Inter-Annotator agreement scores using the available GATE module revealed the agreement between annotators with respect to the targeted concepts. Based on the resulted F-Measure metric the three experts agree 65% in 'average mode' where partial matches count as half matches and 75% in 'lenient mode' where partial matches are measured as full matches. The overall IAA is comparable to the

[3] Implicit since the text is an OASIS archaeological excavation report

results of Archaeotools project and indicative of the inherited subjectivity in the annotation of cultural heritage text [24]. Work is underway investigating the definition of a commonly agreed 'gold standard' version consisted of fifty summary extracts

The preliminary evaluation task made use of ten summary extracts. The manual annotations used by the GATE Corpus benchmark utility producing the overall scores are shown in Table 1. These preliminary results are generally encouraging. If we compare the most basic form of thesaurus expansion (Synonym) with the conceptual expansion modes, they show a slight improvement in F-Measure over all annotators for both Hyponym and Hypernym expansion modes (with larger improvement for Recall). In addition, results show that the system produces better F-Measure scores (two out of three) when performs on the Hyponym expansion mode. However, the F-Measure scores of the Hypernym mode do not differ significantly from those of Hyponym hence the system can also operate on an expansion mode which is in favour of recall than precision. Subsequent and larger scale evaluation will examine further the above trend and will consider whether the F-Measure scores continue to show a similar pattern.

TABLE I.
PRECISION, RECALL AND F-MEASURE SCORES

	Synonym	**Hyponym**	**Hypernym**
Annotator 1			
Precision	0.82	0.85	0.76
Recall	0.67	0.72	0.77
F-Measure	0.73	0.78	0.76
Annotator 2			
Precision	0.72	0.72	0.7
Recall	0.6	0.62	0.68
F-Measure	0.65	0.66	0.68
Annotator 3			
Precision	0.61	0.62	0.6
Recall	0.66	0.69	0.72
F-Measure	0.6	0.62	0.61

VI. CONCLUSION

The discussion has revealed the viability of automatic generation of rich metadata for enabling semantic search of grey literature connected with archaeological datasets. The methods of Information Extraction, driven by the core ontology CIDOC CRM and its extension CRM EH , in combination with SKOS resources, were central to the process of automatic metadata generation. The evaluation results are encouraging and reveal the potential of the method in annotating grey literature documents with respect to CRM while maintaining semantic links to terminological SKOS resources. A large scale evaluation exercise is planned to evaluate the information extraction performance in general and consider lexical ambiguities and the accuracy of rich phrase annotation in particular.

Specific contributions of the research include techniques for automatic rich metadata generation as CRM-EH entities, expression as coupled XML and as RDF triples, cross search over datasets and grey literature, techniques for using SKOS and CRM resources within GATE. In general, the current study demonstrates the capability for CRM based methods to drive automatic generation of rich metadata in domain specific digital libraries. Such metadata can be expressed in interoperable formats such as XML and RDF graphs, which can be exploited by digital library systems to enable cross-search functionality between disparate resources.

ACKNOWLEDGMENT

The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). Thanks are due to the Archaeological Data Service for provision of the OASIS corpus and Phil Carlisle from English Heritage for providing domain thesauri

REFERENCES

[1] Gradmann S: Knowledge = Information in context: on the importance of semantic contextualisation in Europeana. White Paper, http://version1.europeana.eu/web/europeana-project/whitepapers (2010)
[2] Doerr M, Gradmann S, Hennicke S, Isaac A, Meghini C, Sompel van de H : The Europeana Data Model (EDM). In World Library and Information Congress: 76th IFLA General Conference and Assembly, Gothenburg (2010)
[3] Crofts N, Doerr M, Gill T, Stead S, Stiff M, Definition of the CIDOC Conceptual Reference Model. http://www.cidoc-crm.org/docs/cidoc_crm_version_5.0.2.pdf
[4] Doerr, M.: The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 2493, 75--92 (2003)
[5] Cripps P, Greenhalgh A, Fellows D, May K, Robinson D.: Ontological Modelling of the work of the Centre for Archaeology. (2004) <http://www.cidoc-crm.org/docs/Ontological_Modelling_Project_Report >
[6] Online AccesS to the Index of archaeological investigationS (OASIS) at http://www.oasis.ac.uk/
[7] 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), Budapest.
[8] 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, Proc. First International Workshop SIEDL 2008, pp. 88–99, associated with 5th European Semantic Web Conference, Tenerife (2008)
[9] Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proc. 40th Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, 2002
[10] Bontcheva K., Cunningham H., Kiryakov A., Tablan V. Semantic Annotation and Human Language Technology. In Semantic Web Technology: Trends and Research in Ontology Based Systems John Wiley and Sons Ltd (2006)
[11] Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E., and 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 4(1). pp.14–28 (2006)
[12] Reeve, L., and Han H. Survey of semantic annotation platforms. Proceedings of the 2005 ACM symposium on Applied computing. Santa Fe, New Mexico, ACM Press, pp. 1634-1638 (2005)
[13] Popov B, Kiryakov A, Ognyanoff D, Manov D, Kirilov A.: KIM ; a semantic platform for information extraction and retrieval. Natural Language Engineering 10(3-4) pp.375–392. Cambridge University Press, New York (2004)
[14] Jeffrey S, Richards J, Giravegna F, Waller S, Chapman S, and Zhang Z.: The Archaeotools project: faceted classification and natural language

processing in an archeological context. Phil. Trans. R. Soc. A(367) pp 2507–2519. Royal Society Publishing, London (2009)
[15] Ore C-E., Eide Ø.: TEI and cultural heritage ontologies: Exchange of information? Literary and Linguist Computing, 24 (2), 161-172. Oxford University Press (2009)
[16] Isaac A., Summers E.: SKOS Simple Knowledge Organization System Primer, http://www.w3.org/TR/skos-primer (2009)
[17] 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, 62 (4&5), 466 – 475 (2010)
[18] National Monuments Record Thesauri. English Heritage http://thesaurus.english-heritage.org.uk/ thesaurus.english-heritage.org
[19] 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, 280–290. Lecture Notes in Computer Science, 5173, Berlin: Springer (2008)
[20] Chapman W., Bridewell W, Hanbury P, Cooper G., Buchanan B. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics 34(1) pp.301-310, Elsevier Science (2001)
[21] Semantic Technologies for Archaeological Resources (STAR) demonstrator. University of Glamorgan. http://hypermedia.research.glam.ac.uk/resources/star-demonstrator/
[22] Andronikos web-portal of semantic indices of the OASIS corpus (ADS). http://andronikos.kyklos.co.uk
[23] Binding C, May K, Souza R, Tudhope D, Vlachidis A. Semantic Technologies for Archaeology Resources: Results from the STAR Project, Computer Applications and Quantitative Methods in Archaeology (CAA2010), Granada, (2010)
[24] Zhang Z., Chapman S., Ciravegna F. A Methodology towards Effective and Efficient Manual Document Annotation: Addressing Annotator Discrepancy and Annotation Quality. Lecture Notes in Computer Science 6317 pp301–315 (2010) 2.2-6.