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

Andreas Vlachidis[1], Ceri Binding[1], Keith May[2], Douglas Tudhope[1]
[1] Hypermedia Research Unit, University of Glamorgan, UK
[2] English Heritage, UK
{avlachid, cbinding, dstudhope}@glam.ac.uk,
keith.may@english-heritage.org.uk

Abstract. This paper discusses the automatic generation of rich metadata for semantic search of grey literature connected with archaeological datasets. The work is part of the STAR project, in collaboration with English Heritage. An extension of the CIDOC CRM for the archaeological domain acts as a core ontology. This enables cross search of various datasets and an extract of the Archaeological Data Service OASIS library of excavation reports. Rich metadata is automatically extracted from grey literature, directed by the CRM, via a three phase process of semantic enrichment employing the GATE toolkit. This is expressed as XML annotations coupled with reports and RDF metadata,. both expressed as CRM entities, qualified by SKOS archaeological concepts. Examples from two applications are discussed. The Andronikos web portal delivers the annotated XML files for visual inspection. The STAR research demonstrator offers unified search of excavation data and grey literature in terms of the core ontology.

Keywords: Automatic Metadata Generation, CIDOC CRM, Digital Archaeology, Digital Library, GATE, Knowledge Organization Systems, Information Extraction, Semantic Annotation, Semantic Search, SKOS

1 Introduction

It is said that we live in the Information Society. Within this ever increasing complexity of electronic environments, Digital Libraries represent dynamic tools that mediate information, facilitate communication and support interaction between scientists, researchers and the general public. According to the DELOS Network of Excellence on Digital Libraries; A Digital Library is "*a possibly virtual organisation that comprehensively collects, manages and preserves for the long term rich digital content, and offers to its user communities specialised functionality on that content, of measurable quality and according to codified policies*" [1]. Digital Libraries have matured to complex multi-tier architectures, enabling specialised user functionality, sophisticated administration and advanced interoperability management. The increased interest in semantic technologies which has followed the Semantic Web

initiative has brought the semantic paradigm to the forefront of Digital Libraries. Today we witness a shift of Digital Library development towards the potential of semantic contextualisation, employing conceptual models and sharing as much semantic context as possible via open data architectures.

1.1 Semantics for Digital Libraries

A leading example of the shift of digital libraries towards semantic contextualisation is the Europeana project. Described as a digital library, the project links more than 6 million digital items from the cultural and heritage domain [2]. Europeana can be understood as a common ground, an aggregation mechanism for linking digital objects of culture and heritage domain. Although, it offers the functionality of a web portal, Europeana is also defined as an Application Programme Interface (API) on which portal services can be built upon. Importantly, the project delivers a significant semantic enrichment to its linked digital objects. Aiming to enable "complex semantic operations" on the linked resources that would not be possible to deliver by traditional digital library environments, Europeana employs the synaptic Europeana Data Model (EDM) that brings together qualities from a set of well-established conceptual, terminological, and metadata models [3]. The EDM subsumes the CIDOC Conceptual Reference Model (CRM).

1.2 Semantics in the Culture and Heritage Domain

The CIDOC CRM is an international standard (ISO21127:2006) semantic framework, aiming to promote shared understanding of cultural heritage information [4, 5]. The CRM is capable of mapping any type of cultural heritage information, as published by museums, libraries and archives [6]. Extensibility is an important aspect; the CRM can be specialized when it is required by a domain. 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 that addresses the needs for semantic interoperability in the archaeology domain is the English Heritage (EH) extension. The extended CRM-EH model, comprises 125 extension sub-classes and 4 extension sub-properties. Based on the archaeological notion of context, modelled as place, the CRM-EH ontology describes entities and relationships relating to a series of archaeological events, including stratigraphic relationships, phasing information, finds recording and environmental sampling [7]. Thus, Context is a specialisation of Place, ContextFind of Physical Object, ContextEvent of Event, etc.

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 markup [8]. 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.

1.3 Aims and Overview

In archaeology today, we see digital libraries of grey literature reports and of excavation dataset but they are not meaningfully connected. This paper discusses the automatic generation of rich metadata that makes possible semantic search of grey literature connected with diverse archaeological datasets.

The CRM-EH has been used as a core ontology providing a contextual framework between different types of information sources and disparate datasets, by the Semantic Technologies for Archaeological Resources (STAR) project [9, 10]. In collaboration with English Heritage, STAR has developed methods for linking digital archive databases, vocabularies and unpublished excavation reports for purposes of semantic cross search. The CRM-EH is necessary for expressing the semantics and complexities of relationships between the data elements and annotations, which underline semantically defined user queries. The project has also employed knowledge resources, such as EH domain glossaries and thesauri, expressed as SKOS vocabularies [11]. These knowledge resources assist semantically defined queries and NLP information extraction from excavation reports.

An extract from the OASIS (Online AccesS to the Index of archaeological investigations) grey literature library, provided by the Archaeological Data Service, forms the STAR free text corpus. The term, grey literature, is used to describe documents and source materials that cannot be found through the conventional means of publication - many excavation reports exist in this format. The OASIS project is a joint effort of UK archaeology research groups, institutions, and organisations, coordinated by the University of York [12]. It aims to improve the communication of fieldwork results to the wider archaeological community.

This paper focuses on the methods developed for automatically extracting rich metadata from grey literature, directed by the CRM, via a three phase process of semantic enrichment employing the General Architecture for Text Engineering (GATE) toolkit [13]. 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 XML coupled with the grey literature report and also to RDF metadata. Both are expressed as CRM entities, qualified by SKOS archaeological vocabulary concepts. 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. Examples of the semantically enriched grey literature from both applications are discussed in this paper.

2 Semantic Enrichment of Grey Literature

Information Extraction (IE) is a particular NLP technique relevant to the semantic enrichment of free text documents by extracting specific information snippets suitable for further manipulation [14, 15]. These semantic annotations in context enrich documents, enabling access on the basis of a conceptual structure, providing smooth traversal between unstructured text and conceptual models [16]. In addition, they can aid the integration of heterogeneous data sources by exploiting a conceptual structure and allowing users to search across resources for entities and relations, instead of words. Users can search for the term *'Paris'* and a semantic annotation mechanism can relate the term with the abstract concept of *'city'*, while providing a link to the term *'France'*, which relates to the abstract concept *'country'*. Employing a different conceptual model, the same term *'Paris'* can be related with the concept of *'mythical hero'* linked with the city of *'Troy'* from Homer's epic poem, the Iliad.

3 The Process of Semantic Enrichment

The discussion on the process of Semantic Enrichment is divided into the three broad phases of the IE pipeline, each subdivided into various sub-tasks (pipelines). The initial pre-processing phase prepares the OASIS corpus and knowledge resources. The section on the second main IE phase highlights some details of the pipeline used for the annotation of the textual resources with conceptual and terminological references. The last section discusses the techniques employed in the final phase that constructs RDF representations of the metadata for the semantically enriched documents. The discussion begins by introducing the underlying Language Engineering architecture and Knowledge Organization System resources that contribute to the process of semantic enrichment.

3.1 Underlying Architecture

A popular open source Language Engineering platform that can accommodate the task of IE is the General Architecture of Text Engineering (GATE). Developed by the University of Sheffield, GATE is described as an infrastructure for processing human language, a framework and a development environment for developing and deploying natural language software components [13]. The architecture integrates the Java Annotation Pattern Engine (JAPE), enabling the construction of regular expressions in the form of JAPE rules. Rules are synthesised in a cascading order (pipeline) for extracting textual snippets that conform to particular pattern matching rules. In addition, 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).

The open source orientation of the architecture allows modification of ANNIE language resources and integration of a wide range of language processing utilities, distributed in the form of GATE Plug-ins. GATE also enables use of integrated tools for working with conceptual models (OWL Lite Ontologies) and Gazetteers. It allows the processing of a range of different document formats, including Adobe PDF, Microsoft Word and plain text but without maintaining the morphological aspects of the imported documents such as font size and type.

3.2 Underlying 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, and often support named entity recognition tasks. GATE gazetteers are not flat, which means that enlisted terms can enjoy attributes which in turn can be invoked by JAPE rules for the construction of sophisticated patterns. For example, a gazetteer containing the names of European cities can be enhanced with an attribute denoting the country of origin for each European city enlisted in the gazetteer resource. Therefore, a JAPE rule at later stages can exploit that particular attribute for targeting term matching only at UK cities or only at UK and Greek cities. In the semantic enrichment work described in this paper, gazetteers proved useful, being used to accommodate the wide range of Knowledge Organization System (KOS) resources made available to the STAR project by English Heritage.

One type of KOS used in the semantic enrichment process was the thesaurus, with its various semantic relationships. In STAR, the various EH thesauri and glossaries were represented in SKOS, allowing unique identifiers (URIs) for a concept and links between concepts (skos:exactMatch, skos:closerMatch) [11]. Four thesauri and five glossary resources are incorporated in the process of semantic enrichment for identifying occurrences of various conceptual entities. The thesauri are MDA Object Types, Monument Types, Main Building Materials and the Time-line Thesaurus [17]. The glossary resources used in the process are Simple Names for Deposits and Cuts, Find Type Index, Material Index, Small Finds and the Bulk Find Material glossary. All the KOS resources had been previously expressed in SKOS format for the purposes of the STAR project [18, 19].

The glossary resources contain a small set of concepts which are highly relevant to the domain of archaeology. On the other hand, the thesauri resources contain a large set of concepts which relate to the general cultural and heritage domain. Previously, GATE applications had been confined to employing only glossary concepts, which would limit the semantic enrichment to a small set of highly relevant concepts in the STAR context. Thus, new GATE techniques were developed for purposes of the research, which allowed the possibility of exploiting the wider context of the EH thesauri in particular situations. This affords a richer terminology for the semantic enrichment process.

Exploiting the whole range of thesauri resources would expand the process of enrichment to concepts that are not very relevant to the archaeology domain.

Therefore an optimum range of thesauri concepts is needed. A solution is to use concepts that come from those areas of thesauri structures that can be useful to the enrichment process. Overlapping concepts between glossary and thesauri can serve as entry points to the thesauri structures. Thesaurus semantic relationships can then be exploited for expanding from the entry point across thesauri areas relevant to the task of semantic enrichment. In order to enable this semantic expansion within GATE, JAPE rules must be able to exploit narrower and broader semantic relationships of thesauri structures. Therefore, transformation of SKOS thesauri to GATE gazetteers allows the translation of thesauri properties to gazetteer attributes, enabling JAPE rules to exploit semantic relationships between gazetteer terms.

3.3 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.

3.3.1 Transforming KOS to GATE gazetteers

The transformation of SKOS thesauri and glossary resources to GATE gazetteers is 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, which can expand beyond the limits of a single semantic layer. Since JAPE is essentially a pattern matching rule engine, some work was required to enable the semantic expansion of SKOS concepts (with their unique identifiers) within GATE at the lookup stage of the information extraction pipeline.

Consider the following case; *Container (by function) >Food and Drink Serving Container > Drink Serving Container > Jug > Knight Jug*. A JAPE rule that used only the narrower concept property would be able to semantically expand only on 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
Jug@skosConcept=101601@path=/101204/101340/101023
Drink Serving
Container@skosConcept=101204@path=/101340/101023
Food and Drink Serving
Container@skosConcept=101340@path=/101023
Container@skosConcept=101023@path=/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 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 the 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:

```
hearth@skosConcept=ehg003.37@skosExactMatch=70374
```

This allows the potential for JAPE rules to optionally expand the concept Hearth to associated concepts within the Monuments thesaurus, depending on the overall context.

3.3.2 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.

As discussed in the Introduction, the summary is considered an important document section where information extraction of semantic metadata can be considered to have some particular relevance for the document as a whole. The pro-processing phase was able to identify summary sections of the OASIS corpus.

On the other hand, sections such as headings and table of contents are currently considered less relevant for the semantic enrichment process because such sections tend to refer to domain entities abstractly, without any richer discussion. In addition, tabular data are also excluded from the current semantic enrichment process since they do not provide any discussion and 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. The scope of this paper is not to discuss the details of those heuristic patterns. 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.

Last but not least, the pre-processing phase uses ANNIE modules to identify noun and verb phrases of the document. 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’.

3.4 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. The IE process is carried out by the OPTIMA pipeline developed for the

project (Object, Place, TIme and MAterial). These four concepts were considered key metadata elements for the purposes of 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. We term this particular IE approach as Knowledge Based Information Extraction (KBIE), due to the combination of the core ontology and participating knowledge resources playing a major role in the process of semantic enrichment. This drives the IE task by using JAPE patterns which exploit and produce semantic relationships. 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 is to invoke the gazetteer resources and to generate the initial lookup annotations. The various knowledge resources that participate in the pipeline are capable of identifying lookups that fall within the four concept categories mentioned above. For example, the MDA Object Type thesaurus contains concepts of physical objects, the Main Building Materials thesaurus contains concepts of materials, the Timeline thesaurus contains time appellations, such as periods, while glossaries are a source of more specific archaeological terminology, such as the Simple Names for Deposits and Cuts, which covers archaeological contexts. Archaeological contexts in the CRM-EH are modelled as a specialization of Places in the CRM. There are cases where a term can be found within two different knowledge resources, thus potentially having two different conceptual references. For example, the term brick can refer to a material or to a physical object, as can terms such as, glass, stone, iron, gold etc. in the particular domain practice of archaeology. 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 for purposes of the current project. 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 that have to be suppressed by the validation stage. 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]. The NegEx algorithm is modified here 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, thus producing semantic annotations. In order to accomplish annotation at the CRM-EH semantic level, the pipeline invokes a set of

JAPE patterns that finds rich phrases connecting the previously identified lookup entities. The construction of JAPE patterns is informed by a bottom up analysis of the corpus, which has identified the commonly occurring patterns that connect the four different types of entities. These common patterns reveal a Bradford distribution. Following the 80-20 principle, the method used the top 20% of the most frequently occurring patterns to inform the construction of JAPE patterns. The patterns extract phrases that connect lookup annotations of the following CRM-EH entities and re-annotate previously identified CRM annotations to CRM-EH equivalent extension entities:

- *ContextEvent* connecting Place with Time Appellation annotations which are re-annotated with the CRM-EH extensions. *Context* and *ContextEventTimeSpanAppellation*
- *ContextFindProductionEvent* connecting *Physical Object* with *Time Appellation* annotations which are re-annotated as extensions. *ContextFind* and *ProductionEventTimeSpanAppellation*
- *ContextFindDepositionEvent* for connecting *Physical Object* with *Place* annotations which are re-annotated as *ContextFind* and *Context*
- *consists_of* for connecting *Physical Object* with *Material* annotations which are re-annotated as *ContextFind* and *ContextFindMaterial*

The pipeline is capable of exploiting the semantic relationships of the knowledge resources that are accommodated as gazetteer resources. The pipeline can be configured in one of three different modes of semantic expansion developed for purposes of the research within GATE, *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, the last mode of expansion will expand the semantic enrichment task to include the largest set of concepts, while the first mode of expansion will include the smallest and also the most precise set of concepts. For example, for the glossary concept enclosure, Synonym expansion will include the concept garth, in addition to enclosure. *Hyponym* expansion will also include the concepts curvilinear enclosure, ditched enclosure, rectilinear enclosure, etc. and their narrower concepts, such as oval enclosure, double ditched enclosure, polygonal enclosure etc.; *Hyperym* expansion will also include all Monument by Form concepts, such as arch, boundary, barrier, ditch and their narrower concepts, such as fence, hedge boundary ditch etc. Clearly, there are recall/precision trade-offs associated with the different expansion modes and these are a topic for investigation in the forthcoming evaluation work.

3.5 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. The exported XML tags enjoy properties produced during the IE process, such as skos:Concept and skos:exactMatch unique terminological identifier, gateID unique annotation identifier and a note for capturing the context surrounding a semantic annotation. In addition, 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*[1]) modeling the association asserted for data items (mapped to CRM) and SKOS concepts [10]. For example a semantic annotation of the CRM-EH class Context 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 decision to use a server-side scripting language like PHP is supported by the two main requirements, a MySQL database for retrieving the unique file name of each document and a visual interface for parameterisation of the transformations, allowing easy selection of semantic annotations that participate in the transformation. PHP allowed rapid development and proved robust with a large set of documents. The final RDF documents are decoupled from the content. However, as explained above each annotation resource is tied to a corpus wide unique identifier.

The following example presents the details of the RDF transformation. Consider the following rich phrase “pits were uniformly filled with large quantities of pottery”. The phrase can be modeled by a CRM-EH ContextFindDepositionEvent, connecting the find, pottery, with the context, pit. The coupled XML output would have the following structure (the note property is omitted and URIs truncated for simplicity):

```
<EHE1004.ContextFindDepositionEvent
gate:gateId="281105">
 <EHE0007.Context gate:gateId="281155"
skos:Concept="#ehg003.55">pits</EHE0007.Context>
were uniformly filled with large quantities of
 <EHE0009.ContextFind gate:gateId="281158"
skos="#ehg027.2">pottery</EHE0009.ContextFind>
</EHE1004.ContextFindDepositionEvent>
```

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

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>
```

4 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 [18], 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 outcomes, it is seen as indicative of potential digital library applications where access to the semantically enriched text is desired.

The metadata take account of lexical ambiguities such as polysemy (same word having multiple meanings). For example, find all archaeological Contexts of type *"Cut"*, where the term *"cut"* is ambiguous. The semantic enrichment mechanism manages to disambiguate the verb from the noun form and to reveal phrases which make use of "cut" in a archaeological context, eg *"levelling layers sealed the base of a brick wall cut into layer"*, or *"It measured 0.3m in diameter and 0.2m deep with a circular cut"* and to avoid the annotation of non archaeological context *cut*, such as *"although the current 'cut-off' channel is now 500m"*. In addition, the metadata can take account of a form of polysemous ambiguity that is affected by tropony characteristics. For example, the word *"brick"* in an archaeology context can refer to

a material or to a physical object. In the phrase *"yellowish-brown sandy deposit containing frequent unbonded brick"*, the term refers to a physical object, whereas in the phrase *"A layer of small brick tiles forming the street paving"*, the term refers to a material.

The STAR Demonstrator makes use of the rich metadata for some forms of semantic search, building on CRM and SKOS unique identifers. For example, searches are possible of the form: Context of type X containing Find of type Y. The two different extracts of screendumps in Figure 1 show a Context of type *"hearth"* containing Context Find of type *"coin"*, together with a Context Find of type *"Animal Remains"* within a Context of type *"pit"*.

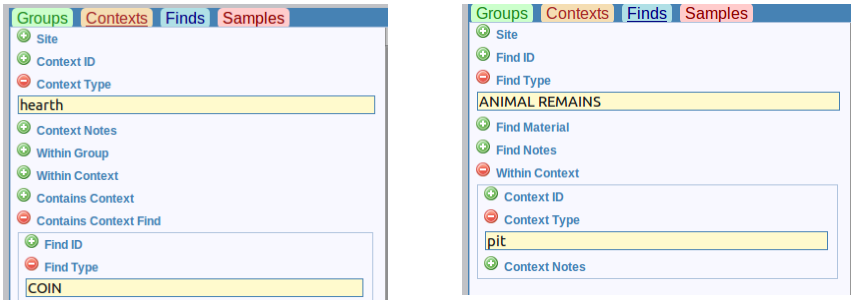

Fig1. The STAR demonstrator search of semantic metadata

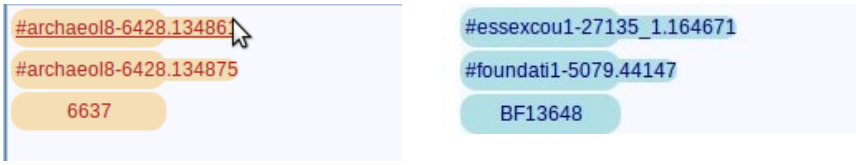

Fig 2. Search Results from the STAR demonstrator (prototype results list user interface)

The cross-search capability of the STAR demonstrator retrieves results from both datasets and grey literature reports (a variety of datasets for the hearth query). As seen above, the searches returned results for different annotation types (Contexts, Finds) and from different resources (grey literature resources commence with a hash bar in this interface). For example, the first search retrieves from Grey Literature #archaeol8-6428.134861 a *Hearth* containing a *coin;* the original text was *"It differs from the other coin finds, however, in that it was associated with a hearth"*. Similarly the second 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 *"Animal Remains"* where the term *"Animal bone"* is retrieved.

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 annotations are divided into three abstractions: (i) preprocessing annotations, such as Headings, Table of Contents and Summary; (ii) single CRM

annotations such as Physical Object, Place, Time and Material; (iii) CRM-EH archaeology specific annotations of rich phrases. Therefore, it is possible to optionally expose particular document abstractions according to different application strategies. Thus, in certain cases, the Summary sections (an example is given in Figure 3) 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 4) 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 annotations (Figure 5). Andronikos also makes available links to the XML and RDF versions of grey literature documents which can be downloaded and further transformed or manipulated.

Fig 3. Summary Section of Grey literature, Andronikos web-portal

TERM	SKOS	Count	TERM	SKOS	Count
anglian	#136306	14	fired clay	#ehg027.5	40
medieval	#134745	19	plate	#96797	53
20th century	#134841	19	artefacts	#ehg020.7	55
prehistoric	#134718	46	tile	#ehg027.3	57
roman	#134738	216	pottery	#ehg027.2	81

Fig 4. Frequent CRM entities (Time Appellation – left side, Physical Object right side) in a grey literature report, Andronikos web-portal

Fig 5. CRM-EH rich metadata

The above examples demonstrate the three different levels of abstraction for the same document. The summary section (Figure 3) discusses evidence that relates to the Roman period, while the CRM overview (Figure 4) show the most frequently used SKOS concepts for the CRM entities Time Appellation and Physical Object, in this case *Roman* and *pottery*. The CRM-EH rich metadata (Figure 5) reveals evidence that makes 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 a, archaeological Context (*sand fill*), via the CRM-EH entity Context Find Deposition event[2]. CRM-EH ContextFind 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 also opens the possibility for very precise semantic queries based upon the connection of entities via the CRM events.

6 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 results of Archaeotools project and indicative of the inherited subjectivity in the annotation of cultural heritage text [24]. Work is underway

	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

Table 1. Precision, Recall and fMeasure scores

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

investigating the definition of a commonly agreed ‘gold standard’ version consitsted 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.

7 Conclusions

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. An early pilot evaluation has revealed the potential of the method in annotating grey literature documents with respect to CRM while maintaining semantic links to terminological SKOS resources [23]. A large scale evaluation excersise is planned to evaluate the information extraction performance in general and in dealing with lexical ambiguities and the accuracy of rich phrase annotation in particular.

Specific contributions of the research include techniques for automatic rich metadata generation and expression as coupled XML and as RDF triples, cross search over datasets and grey literature, technqiues for using SKOS and CRM resources within GATE based information extraction. 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. Work is underway investigating generalisation of the methods to related areas in the cultural heritage domain.

References

1. Candela L, Castelli D, Pagano P, Thanos C, Ioannidis Y, Koutrika G, Ross S, Seamus S, Hans-Jörg S, Heiko S: Setting the foundations of digital libraries. The DELOS manifesto. In D-Lib Magazine, 13 (3/4) (2007)
2. 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)
3. 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)
4. 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
5. Doerr, M.: The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 2493, 75--92 (2003)
6. Babeu, A., Bamman, D., Crane, G., Kummer, R., Weaver, G.: Named Entity Identification and Cyberinfrastructure. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07) Budapest, 259-270 (2007)
7. 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 >
8. 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)
9. 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.
10. 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)
11. Isaac A., Summers E.: SKOS Simple Knowledge Organization System Primer, http://www.w3.org/TR/skos-primer (2009)
12. Online AccesS to the Index of archaeological investigationS (OASIS) at http://www.oasis.ac.uk/
13. 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
14. Moens M.: Information Extraction Algorithms and Prospects in a Retrieval Context. Springer, New York (2006)
15. Cowie J, Lehnert W. Information extraction. Communications ACM 39(1):80–91 ACM, New York (1996)
16. Bontcheva K, Duke T, Glover N, Kings I.: Semantic Information Access. In Semantic Web Semantic Web Technology: Trends and Research in Ontology Based Systems, Wiley, Sussex (2006)
17. National Monuments Record Thesauri. English Heritage http://thesaurus.english-heritage.org.uk/ thesaurus.english-heritage.org
18. 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)
19. Binding C. 2010. Implementing archaeological time periods using CIDOC CRM and SKOS. Proceedings 7th Extended Semantic Web Conference, Heraklion, L. Aroyo et al. (Eds.): ESWC 2010, Part I, Lecture Notes in Computer Science, 6088, 273–287, Springer-Verlag Berlin Heidelberg.
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. Vlachidis A.,Tudhope D. Semantic Annotation for Indexing Archaeological Context: A Prototype Development and Evaluation. Metadata and Semantic Research (Aydin Ozturk ed.), Communications in Computer and Information Science (2011)
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)