Classical Art Semantics Information Extraction: CASIE Pilot Project

Andreas Vlachidis[1] , Douglas Tudhope[1],

[1] University of South Wales, Hypermedia Research Unit, Pontypridd Wales, CF37 1DL, UK
{andreas.vlachidis, douglas.tudhope}@southwales.ac.uk

Abstract. The paper discusses the application of Natural Language Processing (NLP) techniques in the context of semantic annotation of classical art text via rule-based Information Extraction (IE) techniques combined with ontological and domain vocabulary input. The CASIE (Classical Art Semantics Information Extraction) was a pilot collaborative project between the Hypermedia Research Unit (University of South Wales) and the Beazley Archive (Oxford University), which aims to automatically extract information about cultural objects from classical art scholarly texts and represent this information in terms of the ISO metadata standard for cultural heritage, the International Council of Museum's CIDOC Conceptual Reference Model (CRM). In total 12 documents (fascicules – high quality catalogues) were processed, originating from the Corpus Vasorum Antiquorum (CVA) collection containing over 350 high quality catalogues of mostly ancient Greek painted pottery, illustrating more than 100,000 vases. The extracted information was expressed in interoperable RDF graphs consistent with the CLAROS project format. The role of CIDOC-CRM is central for enabling semantic interoperability across the range of datasets that contribute to CLAROS. The CASIE pilot enabled a complementary exploitation of terminological and ontological resources via rule-based information extraction techniques, delivering semantic annotation with respect to the CRM in the broader field of digital humanities

Keywords: Digital Humanities, CIDOC-CRM, Semantic Annotation, Natural Language Processing, Information Extraction.

1 Introduction

In the current web-enabled environment of scholarly endeavours, researchers are increasingly encouraged to make the results of their work available online, both as collections of datasets and document repositories of different kinds. The provision of University research repositories is one topical example and the practice extends to large subject-based collections of research material from major cultural heritage and memory institutions. Attention in now focusing on providing the means to make this online material more easily discoverable and available for reuse and comparative analysis purposes, so that it is not confined to isolated silos of information. This work has been carried out under Digital Library, eScience and Cyberinfrastucture (USA)

banners. It originally focused on scientific and technical information but is now being extended to the Humanities domain. In recent years over seventy, predominantly UK, and European medium/large scale Digital Humanities projects have adopted semantic technologies for data publishing and integration (Isaksen 2011).

Research dedicated to the automatic indexing (annotation) of textual information with controlled keywords is a key aspect for the provision of easily discoverable and interoperable information structures, since human indexing is prohibitively expensive for the volume of data envisaged. A lingua franca is also necessary for effective cross search and comparison; this is achieved by knowledge organization systems and ontologies that provide semantic pathways between concepts and terminology. Thus the automatic extraction of key pieces of information (objects, places, actors and events) in terms of a relevant ontology for the domain will facilitate cross search and discovery (O'Hara et al. 2010).

The paper discusses the prototype development and initial results of the pilot project Classical Art Semantics Information Extraction (CASIE). The project aimed at extracting information about cultural objects from classical art scholarly texts and to represent this information in terms of the metadata standard (ISO 21127:2006) for cultural heritage, the International Council of Museum's CIDOC Conceptual Reference Model (CRM) (Crofts et al. 2009). In particular, information about individual artefacts, in terms of their type (form), dimensions (height-diameter), catalogue reference and description was automatically extracted from a set of 12 fascicules (high quality catalogues) originating from the Corpus Vasorum Antiquorum (CVA) collection. The paper presents the information extraction method of the project and discusses the delivery of interoperable expressions as RDF graphs consistent with the CLAROS project format (Kurtz et al. 2009). It concludes with discussion of the results and recommendations for a future large scale development.

2 Background and Relevant Work

Research efforts aimed at the electronic publishing and dissemination of cultural heritage information with respect to interoperable technologies can be traced back as long as 1985. The Perseus Project (established in 1985, USA) has made a significant contribution to the digital library domain by encoding several thousand documents of early Greek and Latin text (Smith, Ryberg-Cox and Crane 2000). A recent example of the shift of digital libraries towards semantic contextualisation in Europe is the Europeana project, which aggregates more than 6 million digital items from the cultural and heritage domain under a common semantic framework (Gradmann 2010).

A range of research projects aimed at cross-search and retrieval over diverse cultural heritage resources deploy semantic annotation, a particular form of metadata that associates textual instances with ontological definitions, which enables semantic indexing and search facility over disparate resources. Representative examples of this work include: The MultimediaN E-Culture (Schreiber et al. 2010), which deployed Semantic Web technologies for enabling cross-search and retrieval from Dutch cultural heritage collections from the Rijksmuseum Amsterdam and National Museum of Ethnology. The MuseumFinland (Hyvönen et al. 2005), published heterogeneous

museum collections on the semantic web via intelligent content-based search and browsing services. The SINUS project (Staykova et al. 2011), provided a service oriented architecture allowing unified representation and use of heterogeneous resources of Bulgarian iconography. The STAR project (Tudhope et al. 2011), applied semantic and knowledge-based technologies to the digital archaeology domain for developing new methods for linking digital archive databases, vocabularies and the associated grey literature.

The Classical Art Research Online Services (CLAROS) is an international interdisciplinary research initiative led by the University of Oxford, aimed at the semantic integration of world classical art records located in major collections of university research institutes and museums (Kurtz et al. 2009). The project delivers scholarly databases of classical antiquity via a searchable semantic web interface (www.clarosnet.org). The participating collections originate from a wide range of data providers including the Beazley Archive, the German Archaeological Institute, the Ashmolean Museum, the Eastern Art, Jameel Collection, the National Archaeological Museum of Greece, the Benaki Museum and other. The role of CIDOC-CRM ontology is central to enable semantic interoperability across the range of datasets that contribute to the CLAROS infrastructure. It provides a semantic layer which allows representation of the participating data elements with common key concepts while it supports extensibility to future data contributions to the CLAROS framework.

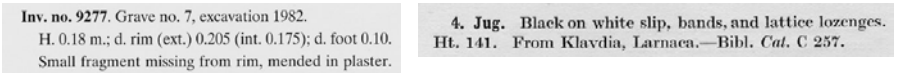
The Corpus Vasorum Antiquorum (CVA) collection is the oldest research project of the Union Académique Internationale organization of national academies in the fields of the humanities and social sciences. CVA was initiated in 1922 and by 2004 has published more than 300 high-quality catalogues (fascicules) of ancient Greek painted pottery illustrating 100,000 vases from over 120 collections located in 26 different countries. In 2004 the Beazley archive completed the digitisation of the CVA fascicules, which resulted in the CVAonline portal (http://www.cvaonline.org). However, the digitisation process is still ongoing since new fascicules are being published.

The CVA digitisation process has delivered text as high quality bitmap images not as machine-encoded text. Although, the content of fascicules is accessible on the Web it cannot be searched or easily lend to a textual operation. It is highly desirable to surface the textual properties of the current bitmap content of CVAonline, in order to enable discoverability of the content of fascicules. An Optical Character Recognition (OCR) process would easily convert the existing bitmap text to machine-encoded output. However, the process would have been incomplete without taking advantage of the current semantic annotation technologies, which can benefit the output of an OCR conversion with conceptual driven metadata that could assign ontological and interoperable characteristics to textual elements. The following sections discuss the method and results of a semantic annotation process over 12 CVA fascicules, which resulted in textual output enhanced with ontological definitions consistent with the CLAROS (CIDOC-CRM) format. Such conceptual definitions of text (semantic annotation) could facilitate cross-searching between CVA fascicules and museum collections that participate in the CLAROS architecture.

3 Method

The semantic annotation process was driven by a rule-based Information Extraction (IE) technique supported by domain-oriented glossaries. Rule-based IE approaches employ hand-crafted regular expressions patterns aimed at recognising relevant textual snippets that are annotated with respect to ontological concepts (classes). The participating glossaries provide relevant domain-oriented vocabulary to the extraction rules, supporting the task of entity recognition with respect to the targeted ontological concepts. Three distinct phases participated in the semantic annotation task; the first phase (pre-processing) performed the OCR task of bitmap text, the second phase (main) implemented the information extraction rules, and the third (export) phase converted the semantic annotations of the second phase to RDF triple graphs consistent with the CLAROS (CIDOC-CRM) format.

Twelve separate fascicules of English text consisting of approximately 500 A4 pages participated in the semantic annotation task. The participating catalogues were published between 1925-1998 and originate from the British Museum, the Ashmolean Museum and the Thessaloniki Archaeological Museum collections. Although they belong to different collections and their publishing date spans from early to late 20th century, their narrative style and structure is reasonably consistent. Earlier catalogues tend to have briefer (2-3 sentences long) descriptions than later publications which offer two to three paragraphs for each artefact discussed (vase or vase fragment). However, on every case of the 12 catalogues, artefact descriptions commence either with a vase shape type (i.e. Amphora, Kotyle, Hydria etc.) or with a catalogue entry (e.g. In.no. 927) as seen in figure 1. This particular consistent narrative of catalogues lends itself to a rule-based IE approach where particular assumptions can be made with regards to the borders of passages that discuss individual artefacts. These can be extracted by uncomplicated regular expressions. On the other hand, specialised rules were formulated for extracting catalogue reference numbers and artefact dimensions which vary in style depending on the origin of individual fascicules.


Inv. no. 9277. Grave no. 7, excavation 1982.
H. 0.18 m.; d. rim (ext.) 0.205 (int. 0.175); d. foot 0.10.
Small fragment missing from rim, mended in plaster.

4. Jug. Black on white slip, bands, and lattice lozenges.
Ht. 141. From Klavdia, Larnaca.—Bibl. *Cat.* C 257.

Figure 1. Example of individual entries in two separate catalogues

The main part of the semantic annotation process was implemented in the natural language processing framework GATE using JAPE rules. The General Architecture for Text Engineering (GATE) is an open source application that provides the architecture and the development environment for developing and deploying natural language software components (Cunningham and Scott 2004). JAPE (Java Annotation Pattern Engine) is a finite state transducer, which uses regular expressions for handling pattern-matching rules (Cunningham, Maynard and Tablan 2000). Such rules are developed and deployed within GATE and enable a cascading mechanism of matching conditions that is usually referred to as the information extraction pipeline. The rules are supported by glossary input that provides domain oriented vocabulary with regards to vase shape types and project specific vocabulary that relates to catalogue and dimension abbreviations. The pre-processing and RDF conversion

phases were implemented outside the GATE environment using commercial applications and bespoke software.

3.1 Pre-processing

The preprocessing phase performed OCR conversion of the original bitmap text to machine-encoded (UTF-8) plain text using the commercial package Abby Fine Reader version 9 (http://www.abbyy.com/). Abby Fine Reader is a popular OCR package known for delivering high quality conversion results. However, the quality of original input in terms of image resolution and contrast is directly linked to the number of erroneous results delivered by an OCR application.

The image resolution of the original bitmap text was high and adequate for conversion to plain text. On the other hand the image contrast was low, including a wide range of grey tones which affect the quality of the conversion result. In order to maximise the performance of the OCR application the image levels of the original bitmap files were adjusted in Photoshop for maximum contrast between white background and black text as seen in figure 2. The high contrast bitmap images were then processed through the OCR application delivering satisfactory conversion results.

Figure 2. Image level adjustment result; left-original image, right-final high contrast result

3.2 Main Information Extraction Phase

The main phase of the semantic annotation process implemented a range of hand-crafted regular expression rules aimed at extracting; i) the descriptive passages of individual artefacts ii) the form type of artefacts, iii) the dimension of artefacts in terms of height and diameter, iv) the catalogue reference number of artefacts. For the extraction of the above, the pipeline employed GATE NLP components, domain-oriented and project specific glossaries and a range of bespoke JAPE rules.

The GATE framework is equipped with a vast range of NLP components (plugins) supporting a number of different general purpose language processing tasks. The CASIE pipeline employed the Tokenizer, the Sentence Splitter and the Part of Speech tagger components of GATE. The NLP components performed the first phase of the pipeline which delivered general purpose annotations, such as tokens equipped with part of speech attributes and sentence sections. The input of the first pipeline phase supports the operation of succeeding bespoke IE rules by providing the necessary annotation input which contributes to the formulation of matching patterns that expressed in rules.

The pipeline also employed the GATE Gazetteer component for providing relevant vocabulary input to the matching patterns. The gazetteer was populated with glossary listings specific to the CASIE project. In particular, three distinct glossary listings contribute to the gazetteer; i) the vase form list which consists of 530 entries of vase shape types, ii) the dimension list which includes abbreviations relating to height and diameter, such as “h.”, “ht.”, “d. rim ext” , “diam” etc., iii) the catalogue reference list which contains catalogue reference abbreviations, such as “Bibl. Cat”, “cat”, “Inv. No” etc. The gazetteer entries are accessible to the IE rules via parameters labelled *Major* and *Minor* type that contribute to the matching expressions.

The bespoke IE rules aimed at recognising descriptive passages, form types and dimension attributes of catalogue artefacts were arranged in a cascading order. The pipeline employed a small set of supplementary matching grammars which supported the operation of the bespoke IE rules. The supplementary grammars were addressed at identifying the starting and end tokens of individual passages which were then used as input to succeeding rules targeted at extracting the descriptive passages of individual artefacts. For example the rule *({VaseTypeHeading}*({EOV}|{EOF}))* matches the description of an individual artefact that commences with a *VaseTypeHeading* annotation type and ends in End of Vase (*EOV*) or End of File (*EOF*) annotation type.

Extraction of textual snippets containing information about the form type of vases employed elaborate rules which combined gazetteer input with regular expressions. For example the rule below matches vase form definitions which are either purely gazetteer entries or contain a gazetteer entry or include a moderator description of a gazetteer entry.

```
{Lookup.majorType==shape}|{Token contains
Lookup.majorType==shape}|
((({Token.category==RB}|{Token.category==NNP}|{Token.ca
tegory==JJ})({SpaceToken.kind==space})?)[1,3]
({SpaceToken.kind==space})*{Lookup.majorType==shape})
```

The above rule will match cases such as, “Amphora”, “Neck-Amphora” and “Fragment of belly of amphora”. Similarly, the rules responsible for extracting dimension and catalogue references of artefacts employ elaborate patterns for matching a range of possible writing styles, such as “Ht. 170”, “Height 170”, “h.170” etc.

3.3 RDF Conversion Phase

The RDF conversion phase exported the semantic annotations from the GATE environment and to express them as interoperable RDF triples consistent with the CLAROS (CIDOC-CRM) format. GATE is equipped with a Flexible Exporter component which exports semantic annotations as XML tags. A bespoke PHP script was developed that used the Document Object Model (DOM) for converting the individual XML tags to RDF triples. Figure 3 presents the graph of RDF triples produced by the PHP conversion script.

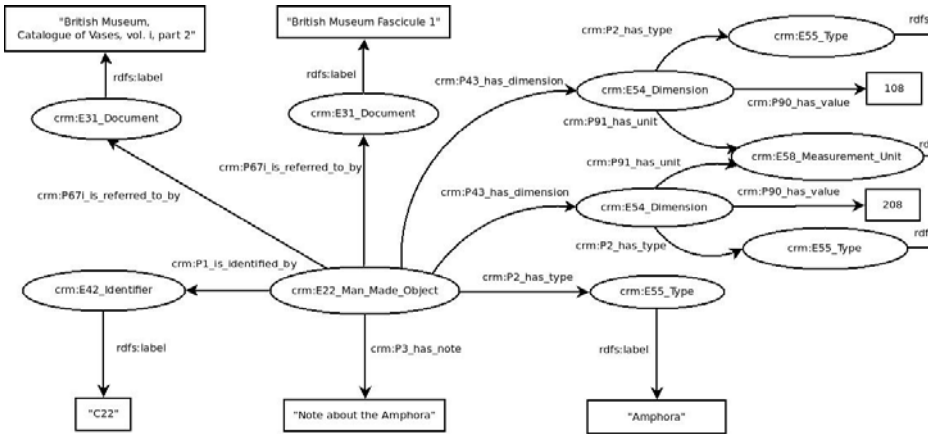


Figure 3. RDF conversion graph of semantic annotation of individual catalogue artefacts

4 Results

Due to the limited resources available to the pilot project, the results were evaluated via visual inspection. According to the preliminary evaluation task, the CASIE pipeline has delivered reasonably good results with regards to the semantic annotation of classical vase descriptions originating from 12 CVA fascicules. Formal evaluation methods aimed at benchmarking the system's performance in terms of Precision and Recall scores fall in the scope of future a full-scale project.

The information extraction task delivered annotations reflecting the descriptive passage, form type, height, diameter and catalogue reference of individual artefacts. Figure 4 presents an example of semantic annotations of 5 individual artefacts (Askos, Lekythos, Small stamnos, Jug with spout and Mastos) as delivered in the GATE environment. The different highlight colours (dark grey in black and white print) represent the semantic annotations of individual vases relating to form type (e.g. Lekythos), height (70) and catalogue reference (C 791). The long highlighted area (light grey in black and white print) reflects the descriptive passage of an individual artefact that commences with form type and ends in catalogue reference.

The RDF conversion script produced triples of semantic annotation as shown in Figure 5. The CIDOC-CRM entity *E22_Man-Made_Object* is used for modelling the individual catalogue artefacts delivered by the CASIE pipeline. The conversion process generated a set of RDF expressions for each individual artefact to accommodate the graph statements of Figure 3. In particular the RDF statements implement the following expressing with respect to a *E22_Man*-Made_Object: i) *P1_is_identified_by* → *E42_Identifier* (catalogue reference e.g. C768) ii) *P2_has_type* → *E55_Type* (artefact type e.g. Jug) iii) *P3_has_note* (the descriptive passage of artefact), iv) *P43_has_dimension* → *E54_Dimension,* which is an abstract node for expressing measurement unit, type and value of artefact width or diameter.

Each fascicule might be composed by several volumes. Therefore, the expression *P67_is_referred_to_by → E31_Document* could be implemented more than once to enable retrieval of an individual artefact either via fascicule or via volume reference.

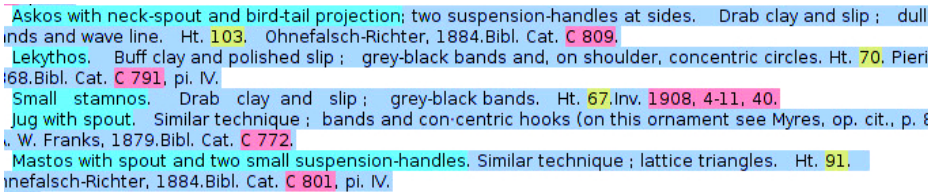

Askos with neck-spout and bird-tail projection; two suspension-handles at sides. Drab clay and slip ; dull
nds and wave line. Ht. 103. Ohnefalsch-Richter, 1884.Bibl. Cat. C 809.
Lekythos. Buff clay and polished slip ; grey-black bands and, on shoulder, concentric circles. Ht. 70. Pieri
68.Bibl. Cat. C 791, pi. IV.
Small stamnos. Drab clay and slip ; grey-black bands. Ht. 67.Inv. 1908, 4-11, 40.
Jug with spout. Similar technique ; bands and con·centric hooks (on this ornament see Myres, op. cit., p.
. W. Franks, 1879.Bibl. Cat. C 772.
Mastos with spout and two small suspension-handles. Similar technique ; lattice triangles. Ht. 91.
nefalsch-Richter, 1884.Bibl. Cat. C 801, pi. IV.

Figure 4. Semantic annotation examples delivered in GATE environment

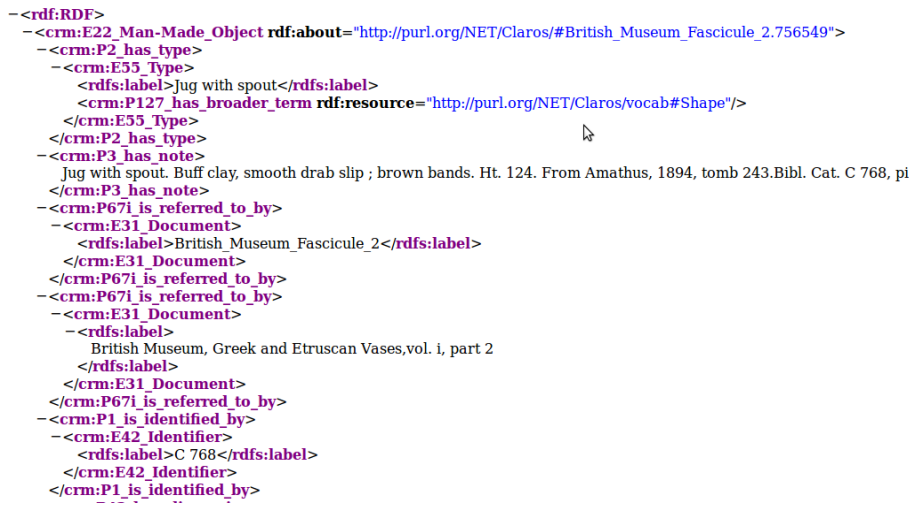

<rdf:RDF>
<crm:E22_Man-Made_Object rdf:about="http://purl.org/NET/Claros/#British_Museum_Fascicule_2.756549">
<crm:P2_has_type>
<crm:E55_Type>
<rdfs:label>Jug with spout</rdfs:label>
<crm:P127_has_broader_term rdf:resource="http://purl.org/NET/Claros/vocab#Shape"/>
</crm:E55_Type>
</crm:P2_has_type>
<crm:P3_has_note>
Jug with spout. Buff clay, smooth drab slip ; brown bands. Ht. 124. From Amathus, 1894, tomb 243.Bibl. Cat. C 768, pi
</crm:P3_has_note>
<crm:P67i_is_referred_to_by>
<crm:E31_Document>
<rdfs:label>British_Museum_Fascicule_2</rdfs:label>
</crm:E31_Document>
</crm:P67i_is_referred_to_by>
<crm:P67i_is_referred_to_by>
<crm:E31_Document>
<rdfs:label>
British Museum, Greek and Etruscan Vases,vol. i, part 2
</rdfs:label>
</crm:E31_Document>
</crm:P67i_is_referred_to_by>
<crm:P1_is_identified_by>
<crm:E42_Identifier>
<rdfs:label>C 768</rdfs:label>
</crm:E42_Identifier>
</crm:P1_is_identified_by>

Figure 5. RDF conversion of semantic annotation consistent with CLAROS format

Conclusion

In recent years there has been a growing attention for the role of semantic technologies in electronic publishing and dissemination of cultural heritage information. The semantic annotation of textual information is a key aspect for the provision of easily discoverable and interoperable information structures, which could enhance current indexing practices (Vlachidis 2012). The paper presented the results of the pilot project CASIE, which aimed at extracting information about cultural objects from classical art scholarly texts. The project delivered a reasonable volume of semantic annotations with respect to form type, dimension, catalogue reference and descriptive information of individual classical objects (vases).

In total 12 CVA fascicules (high quality catalogues) have been processed originating from British Museum, Ashmolean Museum and Thessaloniki Archaeological Museum collections. The produced annotations are consistent with the CLAROS project format, which adopts the metadata standard (ISO 21127:2006) for cultural heritage CIDOC-CRM. The initial results are encouraging and demonstrate the capacity of rule-based information extraction techniques equipped with domain-oriented vocabulary to address semantic annotation over structured catalogues of cultural heritage collections such as the CVA fascicules. The pilot project could form the basis of a future large scale project which could extend and enhance the semantic annotation techniques over the full CVA collection with respect to the multilingual and writing style characteristics of individual fascicules.

Acknowledgments

Thanks are due to the Beazley Archive for providing the CVA fascicules and to the CLAROS team (Oxford University); in particular to Donna Kurtz, Sebastian Rahtz, David Shotton, Graham Klyne and Greg Parker for their valuable input and support.

References

Crofts, N., Doerr, M., Gill, T., Stead, S., and Stiff, M. 2009. *Definition of the CIDOC Conceptual Reference Model, FORTH, Greece*
Available at <http://cidoc.ics.forth.gr/docs/cidoc_crm_version_5.0.1_Mar09.pdf>

Cunningham, H., Scott, D. 2004. Software Architecture for Language Engineering. *Natural Language Engineering.* 10(3-4), 205-209

Cunningham, H., Maynard, D., and Tablan, V. 2000 *JAPE a Java Annotation Patterns Engine* (Second Edition). Technical report CS--00--10, University of Sheffield,
Available at http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0010.pdf

Gradmann S. 2010. Knowledge = Information in context: on the importance of semantic contextualisation in Europeana, *Europeana White Paper*,
Available at <http://version1.europeana.eu/web/europeana-project/whitepapers>

Hyvonen, E., Makela, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M. and Kettula, S. 2005. MuseumFinland – Finnish museums on the semantic web. *Journal of Web Semantics*, 3(2), 224–241

Isaksen L. 2011. *Archaeology and the Semantic Web*. PhD Thesis: University of Southampton. Available at: <http://eprints.soton.ac.uk/206421/ >

Kurtz, D., Parker, G., Shotton, D., Klyne, G., Schroff, F., Zisserman, A., and Wilks, Y. 2009. Claros - bringing classical art to a global public. In *Fifth IEEE International Conference on e-Science.* Oxford, UK

O'Hara K., Berners-Lee, T., Hall, W., Shadbolt, N. 2010. *Use of the Semantic Web in e-Research*. Cambridge: The MIT Press. p131.

Schreiber, G., Amin, A., Aroyo, L., Van Assem, M. 2008 Semantic annotation and search of cultural heritage collections: The Multime diaN E-Culture demonstrator. *Web Semantics: Science, Services and Agents on the World Wide Web* 6 (4), 243-249

Smith D. A., Rydberg-Cox J.A., Crane G.R. 2000. The Perseus Project: a Digital Library for the Humanities, *Literary and Linguistic Computing*, 15(1), 15-25

Staykova K., Agre G., Simov K., and Osenova P. 2011. Language Technology Support for Semantic Annotation of Iconographic Descriptions. In: *Proceedings of the International Workshop "Language Technologies for Digital Humanities"*, Hisar Bulgaria.

Tudhope D, May K, Binding C, Vlachidis A. 2011. Connecting archaeological data and grey literature via semantic cross search, *Internet Archaeology*, (30). Available at: <http://intarch.ac.uk/journal/issue30/tudhope_index.html/>

Vlachidis A. 2012. *Semantic Indexing via Knowledge Organization Systems: Applying the CIDOC-CRM to Archaeological Grey Literature*. PhD Thesis, University of South Wales Available at: <http://hypermedia.research.southwales.ac.uk/media/files/documents/2013-07-11/Andreas-Vlachidis_Thesis_print_ready.pdf>