Pattern based mapping and extraction via CIDOC CRM

Douglas Tudhope[1], Ceri Binding[1], Keith May[2], Michael Charno[3]
([1]University of South Wales, [2]English Heritage, [3]Archaeology Data Service)

douglas.tudhope@southwales.ac.uk
ceri.binding@southwales.ac.uk
keith.may@english-heritage.org.uk
michael.charno@york.ac.uk

1 Introduction

The current situation within archaeology is one of fragmented datasets and applications, with different terminology systems. The interpretation of a find may not employ the same terms as the underlying dataset. Searchers from different perspectives may not use the same terminology. Separate datasets employ distinct schema for semantically equivalent information. Entities and relationships may have different names but be semantically equivalent. Even when datasets are made available on the Web, effective cross search is hampered by semantic interoperability issues [1].

It is becoming increasingly understood that the use of an integrating conceptual framework, such as the CIDOC Conceptual Reference Model (CRM) (ISO 21127:2006) [2, 21], can help address these issues. We take this as our agreed point of departure. This paper discusses various implementation issues to facilitate use of the CRM. Employing the CRM has tended to require an understanding of the source dataset schema and also specialist knowledge of the CRM and techniques for mappings. This paper argues for the use of mapping patterns to guide deployment, to improve homogeneity, to increase data interchange and to encourage greater uptake.

1.1 Relevance to CRMEX Workshop

This paper discusses our implementation experience related to the issues raised in the call for papers of the CRMEX Workshop:

- Because CRM allows many different ways of representing the same situation, CRM adopters in various cultural heritage areas need mapping guidelines and best practices to increase the chance of interoperation.
- While Resource Description Framework (RDF) is a viable CRM representation, there are various low level RDF issues that are not standardized. Since RDF representation implies a certain implementation bias and still undergoes changes of good practice, the CRM Special Interest Group (CRM-SIG) has been expecting good practices to emerge from people applying CRM in order to make recommendations.

The work presented here discusses experience with our development of lightweight techniques and tools to map and extract CRM-based archaeological data with final publication as Linked Data. These techniques have been used in significant CRM-based implementations in two projects STAR [6] and STELLAR [7] described below.

At the Workshop on the Implementation of CIDOC-CRM, organised by the German Archaeological Institute (DAI) in Berlin 2009 [8], we raised the following CRM implementation issues from our experience in the STAR project:

- For application interoperability we need agreement on lower level implementation representations (e.g. data types, date formats, spatial coordinates etc.)
- Need provision of vocabulary (terminology) - our approach is to employ SKOS to model vocabulary elements and link to CRM [19]
- CRM can be extended for domain specificity
- CRM is event-based and therefore
  - ─ Mapping a data property to CRM typically results in a chain of CRM relationships
  - ─ Directly representing the model results in complex user interfaces
  - ─ There is a need for user interface 'short cuts' and simplified views for particular purposes
- Data can be mapped to multiple CRM elements depending on what is considered relevant and important - need for guidelines as to the focus and purpose of a mapping exercise

We next describe briefly the STAR and STELLAR projects, where we explored the above issues. This paper focuses mainly on a discussion of

mapping issues (details of our implementations are given elsewhere but we are happy to discuss in the workshop). We then consider issues raised at the 2009 DAI workshop, together with a discussion of the pattern based approach we have adopted as one way of addressing the issues.

2 STAR Project

The STAR (Semantic Technologies for Archaeological Resources) project was a collaboration between the Hypermedia Research Unit at the University of South Wales (formerly Glamorgan) and English Heritage (EH). The project aimed to provide a degree of semantic interoperability between diverse archaeological datasets from different projects and organisations. The system makes cross-search possible on excavation datasets including Raunds Roman, Raunds Prehistoric, Museum of London, Silchester Roman and Stanwick sampling together with archaeological reports extracted from the OASIS grey literature library, provided by the Archaeology Data Service [9].

Since the CRM operates at a relatively high level of generality, the datasets were mapped to the CRM-EH archaeological extension of the CRM, developed by English Heritage [3, 4]. For working with archaeological datasets at a more detailed level, the CRM-EH specializes the CRM classes for Physical Object and Place to archaeological subclasses such as Find and Context. In collaboration with EH, an RDF implementation was created [4], referencing and complementing the existing published (v4.2) RDFS implementation of the CRM [5].

Domain expert May generated a series of spreadsheets showing the key mappings from the various datasets to the CRM-EH. Selections from the different databases were extracted via SQL queries; and converted to RDF using a data extraction and conversion tool [10].

Despite the use of the data extraction tool the exercise proved time consuming. The initial mappings produced were incomplete and under-specified, relating selected data fields to CRM-EH entities but often at a higher level than that required for implementation. The fully formed intermediate chains of events and relationships necessary for connecting the entities together had to be deduced in each case and conventions unilaterally decided for important implementation details, such as formats for identifiers, coordinates and measurement units.

The online STAR demonstrator cross searches excavation datasets from the five different databases, together with metadata representing an extract of excavation reports from the OASIS grey literature library [22]. STAR did not necessarily seek to represent each dataset in its entirety but focused on specific inter-site cross search use cases. Previously cross search was not possible; each dataset remained in its own silo, and no link was made to grey literature. The demonstrator seeks via the user interface to hide the complexity of the underlying ontology, while offering structured semantic search. An interactive query builder offers search (and browsing) for key archaeological concepts such as Samples, Finds, Contexts or interpretive Groups with their properties and relationships. As the user selects via the interface, an underlying semantic query is automatically constructed in terms of the corresponding ontological model.

STAR employed a web service architecture for programmatic access to the data and to various glossaries and thesauri. The latter were represented in the W3C standard Simple Knowledge Organization System (SKOS) format [11], a formal RDF representation. EH thesauri were available for programmatic access via a web service API, with extensions for semantic concept expansion [20]. The web services were accompanied by a variety of 'widget' controls that could be integrated into browser based user interfaces, where browsing of concept structures or concept based search is required. In more recent work, we have published national heritage thesauri as Linked Data [12].

Natural language processing information extraction techniques were applied to identify key concepts in the grey literature, producing semantic metadata in the same CRM-EH based representation as the extracted data. This metadata allowed unified searching of the different datasets and the grey literature in terms of the semantic structure of the CRM-EH ontology [23].

The CRM and CRM-EH do not supply a vocabulary of concepts beyond the class names in the ontology. Therefore a selection of thesauri and glossaries were used in conjunction with the ontology for search purposes. An extended set of EH glossaries were closely identified with associated fields in the datasets. This required an intellectual alignment operation to cleanse and align the data with controlled vocabulary concept identifiers – an important aspect of the work. These vocabularies afforded semantic search in the demonstrator, with controlled terms being interactively suggested by the query builder.

2.1 STELLAR

STAR served as the launching point for STELLAR (Semantic Technologies Enhancing Links and Linked data for Archaeological Resources) [7], a collaboration between the University of South Wales and the Archaeology Data Service, with EH as Project Partners. We addressed the mapping difficulties discussed in Section 2 by developing new STELLAR tools to make the process more standardised and to facilitate use by third-party data providers. The aim was to make it easier for data owners who are not ontology specialists to express their data in terms of the CRM (and CRM- EH) and to generate Linked Data representations. The STELLAR tools convert archaeological data to RDF in a consistent manner without requiring detailed knowledge of the underlying ontology.

These tools work from a set of templates that express commonly occurring patterns encountered in the STAR project. A set of pre-defined templates is provided but user-defined templates can also be created. The current set of templates corresponds to the general aim of cross-searching excavation datasets for inter-site analysis and comparison. Different templates drawing on other areas of the ontology (and the datasets) could be designed for purposes such as project management and workflow or detailed intra-site analysis. Each template input is a combination of various optional fields with a mandatory ID. The ID is prefixed with a namespace (supplied by the user) to generate URIs. Thus the RDF output is produced in a form that facilitates subsequent expression as Linked Data. The STELLAR template-based method can be considered as a form of the *pattern based approach* that has recently emerged within Linked Data generally [18].

In addition to CRM-based templates, there is a template allowing a glossary or thesaurus connected with the dataset to be expressed in SKOS. The CRM templates have fields giving the (preferred) option of expressing controlled data items as URIs (either to local vocabularies generated by the SKOS template, or to external Linked Data URIs).

Figure 1 is an example of a pattern to model the relationships between an object, a production event and a material.

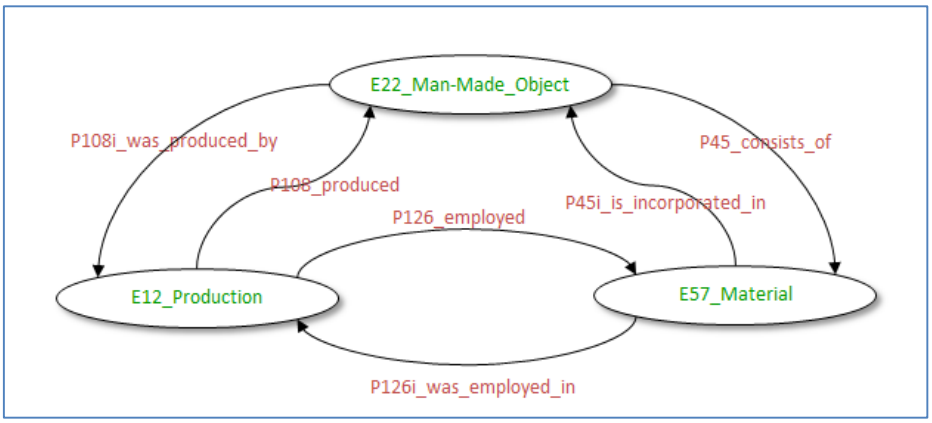

Figure 1. Example pattern

In Figure 2 we see (an extract of) input to the template and then the template itself, which creates directional relationships, an event based property and a shortcut. The user needs to select the particular template (e.g. from a template library) as appropriate for the pattern they wish to express and then supply the data from their datasets. The template contains placeholders corresponding to named columns in the input.

id	material
123	copper

```
// HEADER template, is output once at start
HEADER(options) ::= <<

     <?xml version="1.0"?>
       <rdf:RDF
       xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
       xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
       xmlns:crm="http://www.cidoc-crm.org/cidoc-crm/">

>>
// end of HEADER template

// RECORD template, is output once per data row
RECORD(options, data) ::= <<

  <crm:E22_Man-Made_Object             rdf:about="http://myexam-
ple/E22_$data.id$" />
  <crm:E12_Production rdf:about="http://myexample/E12_$data.id$"
/>
  <crm:E57_Material  rdf:about="http://myexample/E57_$data.mate-
rial$" />
```

```
  <rdf:Description rdf:about="http://myexample/E22_$data.id$">
  <crm:P45_consists_of              rdf:resource="http://myexam-
ple/E57_$data.material$" />
  <crm:P108i_was_produced_by        rdf:resource="http://myexam-
ple/E12_$data.id$" />
  </rdf:Description>

  <rdf:Description   rdf:about="http://myexample/E57_$data.mate-
rial$">
  <crm:P45i_is_incorporated_in      rdf:resource="http://myexam-
ple/E22_$data.id$" />
  <crm:P126i_was_employed_in        rdf:resource="http://myexam-
ple/E12_$data.id$" />
  </rdf:Description>

  <rdf:Description rdf:about="http://myexample/E12_$data.id$">
  <crm:P108_has_produced            rdf:resource="http://myexam-
ple/E22_$data.id$" />
  <crm:P126_employed                rdf:resource="http://myexam-
ple/E57_$data.material$" />
  </rdf:Description>

  >>
  // end of RECORD template

  // FOOTER template, is output once at end
  FOOTER(options) ::== <<
       </rdf:RDF>
  >>
  // end of FOOTER template
```

Figure 2. Example of a STELLAR template and input extract

Templates are available from the STELLAR website, along with tools that operate over the templates. To generate RDF, the user chooses a template for a particular data pattern and supplies the corresponding input from their database. Documentation and a tutorial are available on the website [7]. The Archaeology Data Service used the STELLAR tools to publish Linked Data from a (new) selection of their archived excavation datasets [13].

3 CRM implementation experience from 2009 DAI workshop

Two other projects at the 2009 DAI workshop raised overlapping issues though following different specific implementation methods. The CLAROS project [14] followed a pattern based approach by requiring

data providers to conform to a set of XML format CRM patterns [15]. The BRICKS project discussed below encountered various problematic issues when attempting semantic interoperability via the CRM.

The BRICKS FP6 IP project [16] employed spreadsheets to intellectually define mappings from two different archaeological databases to the CIDOC CRM. These were semi-automatically transformed to XSL style sheets, which transformed the data to the desired representation. They experienced consistency problems which resulted in different mappings for the same underlying semantics and in different data objects being mapped to the same CRM entity. They suggested a need for additional technical specifications for implementation modeling purposes. The abstractness of the CRM and the lengthy relationship chains arising from the event-based model also raised issues for designing appropriate user interfaces.

Further details are elaborated in [17] with various potential opportunities for divergent mappings of the same semantics outlined. Examples are given below (Figure 2 illustrates the first two points):-

- Should an E57 Material (e.g. *gold)* be mapped as a property of an E11 Modification event or as a property of an E22 Man-Made Object?
- Should a method of manufacture (e.g. *hammered*) be mapped as an E55 Type of an E12 Production event or as an Appellation of an E29 Design or Procedure?

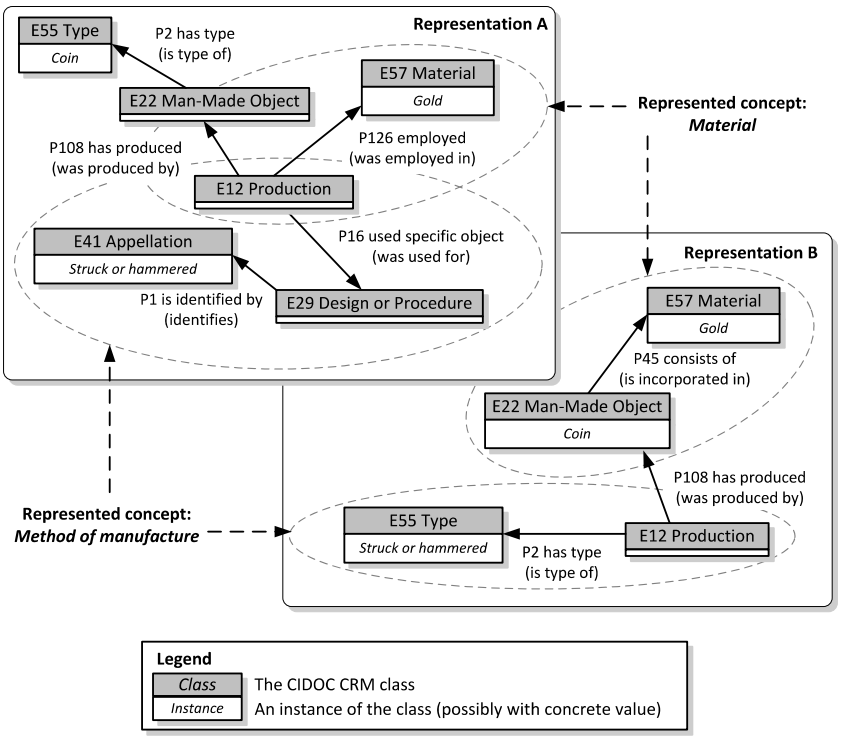

Figure 4: Different valid CRM representations for equal metadata attributes

Figure 2 – a figure taken from [17] illustrating the previous points

Note that the alternatives in Figure 2 are not necessarily equivalent; using a material does not necessarily mean incorporating it in the product and being incorporated does not always imply its use in production. For this instance, both mappings were seen as equally possible in [17] – the note associated with the coin reads "Roman Gold aureus of Nero (AD 54-68) …". Their argument is for more guidance on defining the mapping paths.

- Should E22 Man-Made Objects be directly identified by an E42 Identifier or should the connection be made via a record that has an Identifier? Due to the CRM's origins in museum documentation systems, CRM-based integration work has sometimes modeled the record of an object as an entity in its own right. This can give rise to differences with approaches that seek to directly model an object without noting any existing catalogue or recording element.

- All CRM classes can be assigned types (used for domain terminology). This allows different judgments as to whether a thesaurus or gazetteer element should be associated with an object or related activity (or indeed any property).

In addition to the various mapping choices outlined above we can also note that core ontologies offer the flexibility of capturing different aspects of an object, depending on intellectual judgment. Depending on the end purpose of the mapping exercise, a given aspect may or may not be important to model, as for example perhaps with Man Made Objects and Legal Objects, or man-made features. This will naturally vary between different collections with different areas of focus.

Since the CRM is event-based, the issue of when it is appropriate to create an assignment event when assigning an attribute to an object is ever present. Essentially this depends whether the decision to assign an attribute is considered worthy to record. Is the time and actor involved important? Might others judge differently now or in the future? Again this can result in different mapping expressions depending on the judgement.

It could be argued that the choice to model either a shortcut property or a longer fully formed event-based chain adds flexibility. However, inevitable inconsistencies of approach can result. The STELLAR solution is for the templates to automatically generate a pattern of entities and properties consistently modelling *both* possible approaches simultaneously, thus reducing inconsistencies and the requirements for end applications to detect or predict which particular modelling approach has been taken.

Different mappings can potentially pose significant problems for semantic interoperability. It indeed proved a problem for the BRICKS project, which required the addition of an intermediate mapping which itself served as the integrating layer rather than the CRM. In fact, any general core ontology will permit various mappings from the same set of data elements depending on end purpose and focus.

In principle, end-application systems, capable of intelligently traversing the different CRM graphs produced by differences in mapping practice and differences in the granularity of detail and events modelled, could automatically address the issue of different mappings. In previous work with the Art and Architecture Thesaurus, we have implemented faceted query expansion [24]. With regard to the CRM, Tzompanaki and

Doerr [25] discuss the potential for automatic reasoners to take advantage of transitive properties, propagating down from a query expressed in terms of small set of high level fundamental categories and properties (or offering successive specialised choices to the user). While this offers potential approaches for starting from high level facets, in some use cases the ability to start from lower level query patterns is desirable. The performance issues remain to be fully explored (they point out the deficiencies of SPARQL for such complex queries).

The potential to employ reasoning over the CRM graph is indeed one of the reasons for semantic integration. It defeats the point of integration if everyone must say exactly the same thing with the CRM! Nonetheless in our view, a multiplicity of approaches for similar data will pose unnecessary problems for implementation in the medium term. It is not clear that all the problems described by the BRICKS team could be solved by transitive closure alone. Specific rules will probably be required, which raises difficulties for generalising and introducing a new alternative mapping. A pragmatic approach is to combine developments in reasoning with efforts at consensus on patterns for CRM mappings and guidelines. This could involve patterns for particular domains and also general patterns for common situations.

4 Conclusions

When the CRM was originally created the practical context for automated cross search was more limited and it was in part an intellectual resource. Today there is an expectation that any integrating ontology will be employed in machine readable form for automatic semantic interoperability purposes. However, if different implementations of the CRM follow different low level implementation specifications or employ different mappings for the same underlying semantics then this raises barriers for semantic interoperability.

Issues with mapping are probably inevitable in a general ontology intended to capture a wide range of practice and, as with the application of general library classification schemes, different choices for realising a collection in the CRM may be expected. However the potential divergence of mapping practice poses challenges for implementations and the final applications, particularly where it cannot be assumed that such applications possess built in reasoning capabilities that could ameliorate some of the differences.

Thus the purpose (or use case) of any shared mapping exercise should be stated if possible. Data providers or those responsible for mappings should have available (if they choose) *mapping patterns* and corresponding guidelines for their domain or the mapping exercise in question.

Working from established RDF patterns guarantees the semantic interoperability of the resultant data and also that the syntactical implementation details are handled consistently. It is also more friendly to non-specialists. Mapping patterns were appropriate for the situation with STAR and STELLAR since there was a clear general use case – inter site cross search without requiring clients to possess extensive reasoning capabilities, with the focus on key archaeological concepts [22]. It is possible to define new patterns although this involves more technical expertise.

In some situations there may not be any clear use case that can be reflected in the patterns with which to drive the mapping. Sometimes the use case may emerge following more thorough reflection of the purpose of the mapping exercise. In other situations, it may be considered desirable to capture every aspect of the original dataset for unspecified and unknowable future research purposes. In this case, it may be harder to specify higher level mapping patterns but it should still be possible to specify lower level micro-patterns that can be combined together.

5 Future work

The recent specification by the CRM-Sig of definitive URIs for CRM entities has facilitated one aspect of implementation representation. We need to revise the STELLAR templates and the CRM-EH to conform to this.

We concluded our 2009 DAI workshop presentations with the following proposed issues to take forward, assuming they were considered possible and desirable:

- Agreement on implementation details (e.g. primitives)?
- Agreement on archaeological vocabulary approaches?
- Agreement on archaeological CRM extensions?
- Agreement on mapping patterns and guidelines?

In our view, these issues are still relevant today. We would also add additional aspects – the desirability of expressing the end-purpose of a

mapping exercise; the provision of appropriate registries of mapping patterns; core metadata for mapping patterns together with the means for potential users to discover the patterns.

6 Acknowledgements

The STAR and STELLAR projects were supported by the Arts and Humanities Research Council [grant numbers AH/D001528/1, AH/H037357/1]. Thanks are due to the Archaeology Data Service for provision of the OASIS corpus and for many helpful discussions. Thanks are also due to Phil Carlisle (English Heritage), Andreas Vlachidis (University of South Wales) and the participants of the STAR and STELLAR workshops.

References

1. Patel, M., Koch, T., Doerr, M., Tsinaraki, C.: Report on Semantic Interoperability in Digital Library Systems. DELOS Network of Excellence, WP5 Deliverable D5.3.1. (2005)
2. CIDOC CRM: CIDOC Conceptual Reference Model. Heraklion, Crete: Institute of Computer Science, Foundation for Research and Technology. http://www.cidoc-crm.org/
3. Cripps, P., Greenhalgh, A., Fellows, D., May, K., Robinson, D.: Ontological Modelling of the work of the Centre for Archaeology, CIDOC CRM Technical Paper (2004), http://cidoc.ics.forth.gr/technical_papers.html
4. CRM-EH: English Heritage Extension to CRM for the archaeology domain, http://hypermedia.research.southwales.ac.uk/kos/CRM/ http://purl.org/crmeh
5. RDFS Encoding of the CIDOC CRM, http://cidoc.ics.forth.gr/rdfs/cidoc_v4.2.rdfs
6. STAR Project: Semantic Technologies for Archaeological Resources, http://hypermedia.research.southwales.ac.uk/kos/star/
7. STELLAR Project. Semantic Technologies Enhancing Links and Linked data for Archaeological Resources. University of South Wales: Hypermedia Research Unit. http://hypermedia.research.southwales.ac.uk/kos/stellar/
8. Binding C., Tudhope D. 2009. Breaking Down Barriers to Interoperability. Interconnected Data Worlds: Workshop on the implementation of CIDOC-CRM, organised by the German Archaeological Institute in Berlin and funded by the TOPOI Excellence Project. http://www.dainst.org/medien/de/10_TudhopeBinding_STAR.pdf
9. Archaeology Data Service: Unpublished Fieldwork Reports (Grey Literature Library) http://archaeologydataservice.ac.uk/archives/view/greylit/
10. 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)

11. SKOS: Simple Knowledge Organization Systems - W3C Semantic Web Deployment Working Group http://www.w3.org/2004/02/skos
12. SENESCHAL Project: http://hypermedia.research.southwales.ac.uk/kos/seneschal/
13. Archaeology Data Service: Linked Data. http://data.archaeologydataservice.ac.uk/
14. CLAROS: The world of art on the semantic web. http://www.clarosnet.org/
15. CLAROS Wiki: CIDOC CRM entity description templates for Objects, Places, Periods & People http://www.clarosnet.org/wiki/
16. Nußbaumer, P., Haslhofer, B. (2007). Putting the CIDOC CRM into Practice – Experiences and Challenges. (Technical Report TR-200). University of Vienna. http://cs.univie.ac.at/research/publications/publikation/infpub/404/
17. Nußbaumer, P., Haslhofer, B., Klas W. (2010). Towards Model Implementation Guidelines for the CIDOC Conceptual Reference Model. Technical Report TR-201. University of Vienna. http://eprints.cs.univie.ac.at/58/
18. Dodds, L., Davis, I. (2012). Linked Data Patterns – A pattern catalogue for modelling, publishing and Consuming Linked Data. http://patterns.dataincubator.org/book/
19. Tudhope, D., Binding, C., May, K.: Semantic interoperability issues from a case study in archaeology. In: Stefanos Kollias & Jill Cousins (eds.), Semantic Interoperability in the European Digital Library, Proceedings of the First International Workshop (SIEDL) 2008, 88–99, associated with 5th European Semantic Web Conference, Tenerife (2008)
20. Binding, C., Tudhope, D.: Terminology Services. Knowledge Organization, 37(4), 287–298. (2010). Ergon-Verlag
21. Doerr, M. (2003). The CIDOC conceptual reference model: an ontological approach to semantic interoperability of metadata, AI Magazine 24(3), 75–92
22. Tudhope, D., May, K., Binding, C., Vlachidis, A. (2011). Connecting archaeological data and grey literature via semantic cross search. Internet Archaeology, 30, Open access. http://dx.doi.org/10.11141/ia.30.5
23. Vlachidis, A., Tudhope, D. (2012). A pilot investigation of information extraction in the semantic annotation of archaeological reports. International Journal of Metadata, Semantics and Ontologies, 7(3), 222-235. Inderscience.
24. Tudhope, D., Binding, C., Blocks, D., Cunliffe, D. (2006). Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509-533. Emerald.
25. Tzompanaki, K., Doerr, M. (2012). A New Framework For Querying Semantic Networks. http://www.museumsandtheweb.com/mw2012/papers/a_new_framework_for_querying_semantic_networks Proc. Museums and the Web 2012.