Semantic interoperability issues from a case study in archaeology

D. Tudhope, C. Binding, University of Glamorgan 1

K. May, English Heritage

Abstract. This paper addresses issues arising from the first steps in mapping different (a) datasets and (b) vocabularies to the CIDOC CRM, within an RDF implementation. We first discuss practical implementation issues for mapping datasets to the CRM-EH and then discuss practical issues converting domain thesauri to the SKOS Core standard representation. We finally discuss, at a more theoretical level, issues concerning the mapping of domain thesauri to upper (core) ontologies.

1   Introduction

The general aim of our research is to investigate the potential of semantic terminology tools for improving access to digital archaeology resources, including disparate data sets and associated grey literature. The immediate goal discussed here concerns describing and accessing cultural objects using the CIDOC CRM core ontology [3, 7], as an overarching common schema. Different datasets must be mapped to the CIDOC CRM, where the datasets are indexed by domain thesauri and other vocabularies. Thus semantic interoperability is central.

This paper addresses issues arising from the first steps in mapping different (a) datasets and (b) vocabularies to the CIDOC CRM, within an RDF implementation. The work, in collaboration with English Heritage (EH)[7], formed part of the JPA activities of the DELOS FP6 Network of Excellence, Cluster on Knowledge Extraction and Semantic Interoperability [6] and the AHRC funded project on Semantic Technologies for Archaeological Resources (STAR) [17].

Some previous work in the European DL context (BRICKS) has reported difficulties when mapping different cultural heritage datasets to the CIDOC CRM due to the abstractness of the concepts resulting in consistency problems for the mapping work and also a need for additional technical specifications for CRM implementations [11, 12]. The CRM is a high level conceptual framework, which is intended to be specialised when warranted for particular purposes. We also found a need to provide additional implementation constructs and these are outlined below. For mapping to datasets at a detailed level, we worked with an extension of the CIDOC CRM (the CRM-EH) developed by our collaborators (May) in English Heritage [5, 10]. The CRM-EH models the archaeological excavation and analysis workflow. Working with

---

1 contact address dstudhope@glam.ac.uk

May, an implementation of the CRM-EH has been produced as a modular RDF extension referencing the published (v4.2) RDFS implementation of the CRM. Additional extensions to the CIDOC CRM, necessary to our implementation, are also available as separate RDF files.

We go on to first discuss practical implementation issues for mapping datasets to the CRM-EH and then discuss practical issues converting domain thesauri to the SKOS Core standard representation. We finally discuss, at a more theoretical level, issues concerning the mapping of domain thesauri to upper (core) ontologies.

2 Data extraction and mapping process and conversion to RDF

Initial mappings were made from the CRM-EH to three different database formats, where the data has been extracted to RDF and the mapping expressed as an RDF relationship. The data extraction process involved selected data from the following archaeological datasets:

• Raunds Roman Analytical Database (RRAD)
• Raunds Prehistoric Database (RPRE)
• York Archaeological Trust (YAT) Integrated Archaeological Database (IADB)

The approach taken for the exercise was to extract modular parts of the larger data model from the RRAD, RPRE and IADB databases via SQL queries, and store the data retrieved in a series of RDF files. This allowed data instances to be later selectively combined as required, and avoided the data extraction process from becoming unnecessarily complex and unwieldy.

The intellectual mapping requires some expert knowledge of the data and the CRM-EH. Initial mappings were performed by May and communicated via spreadsheets. Some subsequent mappings were performed by the project team using the initial mappings as a guide, with validation by May. This process is time consuming and a data mapping and extraction utility was developed to assist the process. The utility consists of a form allowing the user to build up a SQL query incorporating selectable consistent URIs representing specific RDF entity and property types (including CRM, CRM-EH, SKOS, Dublin Core and others). The output is an RDF format file, with query parameters saved in XML format for subsequent reuse. Details will be available shortly on the STAR project website.

2.1 ID format adopted

RDF entities require unique identifiers. Some of the data being extracted was an amalgamation of records from separate tables – e.g. *EHE0009.ContextFind* actually contained records from RRAD.Object & RRAD.Ceramics tables. It was therefore necessary to devise a unique ID for all RDF entities beyond just using the record ID

from an individual table.. The format adopted to deal with all these issues was a simple dot delimited notation as follows:

[URI prefix]entity.database.table.column.ID
e.g. "EHE0008.rrad.context.contextno.100999"

This format (although verbose) allowed the use of existing DB record ID values without introducing ambiguities. In RRAD database, Ceramics and Objects were both instances of *EHE0009.ContextFind*. This therefore involved the combination of data from two tables:

• *EHE0009.rrad.object.objectno.105432 [an EHE0009.ContextFind record from the RRAD object table]*
• *EHE0009.rrad.ceramics.ceramicsno.105432 [an EHE0009.ContextFind record from the RRAD Ceramics table, with a coincidental ID value]*

The format also allowed the same base record ID to be used for both *EHE0009.ContextFind* and *EHE1004.ContextFindDepositionEvent* (these records actually originated from the same table and had a 1:1 relationship), using a different entity prefix to disambiguate the records:

• *EHE0009.rrad.object.objectno.105432 [The ContextFind record ID]*
• *EHE1004.rrad.object.objectno.105432 [The ContextFindDepositionEvent record ID]*

Finally an arbitrary URI prefix (http://tempuri/) was added to all ID values. According to need, this can be replaced with a more persistent prefix.

2.2 Date/Time format adopted

There is nothing dictated in CRM or CRM-EH about date/time representation formats, however we clearly needed to maintain a consistent format throughout the data. For the purposes of the data extraction to keep all data consistent we used a "big endian" (i.e. from most to least significant) format compatible with both W3C standards and ISO8601 ("Data elements and interchange formats – Information interchange – Representation of dates and times"). The format is as follows:

CCYY-MM-DDThh:mm:ss e.g. "2007-05-03T16:19:23"

This format does not introduce any restrictions on how dates & times are eventually displayed or used within applications; it merely provides a common string representation mechanism for interoperability of data.

2.3 Co-ordinate format adopted

Spatial co-ordinates appeared in various formats within the datasets. RRAD co-ordinates were 6 digit numeric values in separate "Easting" and "Northing" columns. RPRE coordinates were slash separated string values, sometimes with an extra 4 digit value appended (i.e. either *nnnnnn/nnnnnn/nnnn* or *nnnnnn/nnnnnn*). IADB co-ordinates were numeric values in separate "Easting" and "Northing" columns (and appeared to be relative to a site local reference datum). CRM/CRM-EH requires a single string to represent a spatial co-ordinate value. The consistent format chosen for output was 6 digit space delimited Easting and Northing values, with an optional Height value (Above Ordnance Datum). These values were all assumed to be in metres:

nnnnnnE nnnnnnN [nn.nnnAOD] e.g. "105858E 237435N 125.282AOD"

2.4 Modelling notes/annotations

The CRM has a modelling construct in the form of "properties of properties. For example, property *P3.has_note* has a further property *P3.1.has_type* – intended to model the distinction between different types of note. However, this construct does not translate well to RDF. As evidence of this, property *P3.1.has_type* is not actually part of the current RDFS encoding of CRM on the CIDOC website (in the comment header there is a suggestion to create specific sub properties of *P3.has_note* instead). The more recent OWL encoding of CRM also avoids including the construct.
The EH recording manuals and the current datasets contain several kinds of note fields. Through discussion with EH it is possible to distil these down to a common core set of note types, such as
• Comments
• Method of excavation
• Interpretation
• Siting description
• Site treatment
While it might potentially be restrictive to model notes as strings (notes have other implicit attributes such as language, author/source etc.), this is the current position within the CRM (*E1.CRM Entity _ P3.has_note _ E62.String)*. However, taking the RDFS encoding of CIDOC CRM recommendation, we intend to create sub properties of *P3.has_note* e.g. *EHPxx1.has_interpretation*, as part of future work.

2.5 Modelling of Events

The CRM-EH and CRM are event based models. Events defined in the models and used to interconnect objects and places etc. were often only implicit within the original relational database structures and in the mappings created. In the translation from relational data structures to an RDF graph structure it was necessary to create this event information by the formation of intermediate 'virtual' entities.

2.6 Modelling of Data Instance Values

Being a higher level conceptual model the CRM has little intrinsic provision for the representation of actual data instance values. The approach adopted for the STAR data extraction process was to create rdf:value relationships as an additional property to model instance data for entities wherever appropriate.

2.7 Initial mapping of data fields to extended CRM

The extracted data represented a subset of the full English Heritage extended CRM (CRM-EH) model. For the initial phase we limited the scope of the data extraction work to data concerning contexts and their associated finds. The relationships between entities extracted and modelled in RDF are shown in Figure 1.
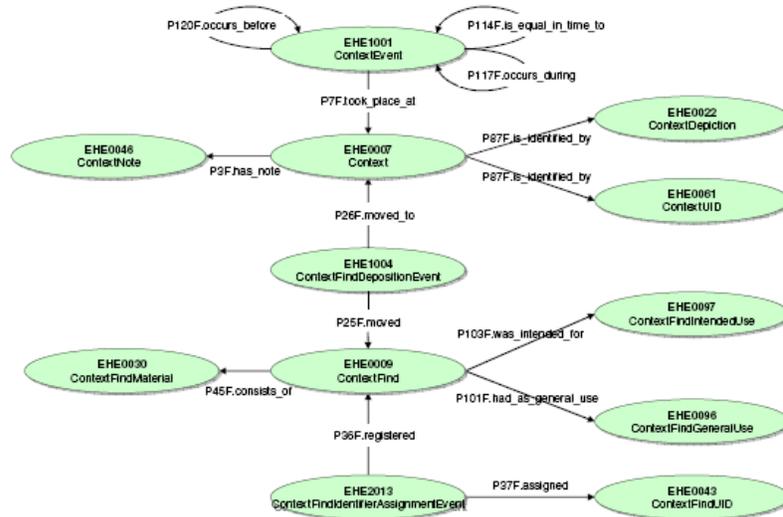
Figure 1: CRM-EH entities initially modelled

The number of statements (triples) contained in the resultant RDF files is **1,080,913.** Some triples (e.g. *rdf:type* statements) were duplicated due to entities occurring within multiple files, but any duplication was removed during the aggregation process.

A number of separate RDF files were combined in the aggregation process including the CRM itself, the CRM-EH extension, alternative language labels for the CRM, and various EH domain thesauri.

2.8 Validation of extracted data

The data files produced were each validated against the W3C RDF validation service. Whilst this did not prove the validity of the data relationships or even conformance to CRM-EH, it did at least give confidence in the validity of the basic RDF syntax.

2.9 Aggregation of extracted data

The SemWeb library [14] was employed to aggregate the extracted data files into a single SQLITE database. The extended CRM ontology plus the English Heritage SKOS thesauri were also imported. The resultant database of aggregated data was 193MB overall and consisted of 268,947 RDF entities, 168,886 RDF literals and 796,227 RDF statements (triples). The SemWeb library supports SPARQL querying against the database, but the SQLITE database itself also supports direct SQL queries.

2.10 Use of aggregated data

This simple, initial example illustrates a SPARQL search via the CRM model relationships for a *Dish* made of *Pewter*. The search is case sensitive and returned 5 records within 1 second. It is possible to deduce the origin of the result records due to the ID convention adopted for the data export process. All are *EHE0009.ContextFind* objects - 3 originated from Raunds Roman (RRAD) *object* table, 1 from the Raunds Prehistoric (RPRE) *flint* table and 1 from the RPRE *objects* table. Merging the exported datasets into the RDF data store facilitates cross searching and location of records from multiple databases.

```
SELECT * WHERE
{
    ?x crm:P103F.was_intended_for "Dish".
    ?x crm:P45F.consists_of "Pewter" .
}
<result>
    <binding name="x">
    <uri>http://tempuri/EHE0009.rrad.object.objectno.12687</uri>
    </binding>
</result>
<result>
    <binding name="x">
    <uri>http://tempuri/EHE0009.rrad.object.objectno.12969</uri>
    </binding>
</result>
<result>
    <binding name="x">
    <uri>http://tempuri/EHE0009.rrad.object.objectno.55006</uri>
    </binding>
</result>
<result>
    <binding name="x">
    <uri>http://tempuri/EHE0009.rpre.flint.recordnumber.55006</uri>
    </binding>
</result>
```

```
<result>
    <binding name="x">
    <uri>http://tempuri/EHE0009.rpre.objects.recordnumber.55006</uri>
    </binding>
</result>
```

3 Conversion of KOS to SKOS/RDF representations

The project has adopted SKOS Core [15] as the representation format for domain thesauri and related Knowledge Organization Systems (KOS). In general, thesauri conforming to the BSI/NISO/ISO standards should map in a fairly straight forward manner to SKOS. However, there may need to be judgments on how to deal with non-standard features. Additionally, the case study illustrates potential problems associated with the use of Guide Terms or facet indicators in some thesauri. Other issues surfaced by the exercise concern the need to create URIs for concept identifiers as part of the conversion and the potential for validation.

3.1 Conversion process

Thesaurus data was received from English Heritage National Monuments Record Centre, in CSV format files [9]. The approach initially adopted was to convert the received files to XML, and an XSL transformation was written to export the data to SKOS RDF format. Although this strategy was successful for the smaller thesauri, XSL transformation of the raw data files proved to be a lengthy and resource intensive operation for the larger thesauri, resulting in the PC running out of memory on some occasions. Therefore the CSV files were subsequently imported into a Microsoft Access database and a small custom C# application was written to export the data from this database into SKOS RDF format.

The major difficulty with the resultant SKOS representations is that we did not model "non-indexing" concepts (guide terms or facet indicators) as *Collections,* the intended equivalent in the SKOS model. Guide terms in SKOS do not form part of the main hierarchical structure, but are groupings of sibling concepts for purposes of clarity in display. It would have entailed changing the existing hierarchical structure of the English Heritage thesauri, in order to utilise the SKOS 'Collections' element. This was not an appropriate decision for the STAR project to take (relevant EH contacts have been informed) and was not a critical issue for the project's research aims. Thus for STAR purposes the distinction between indexing concepts and guide terms is not made, and the (poly) hierarchical relationships in the SKOS files represent those present in the source data.

3.2 Validation process

As a result of running the conversion application, separate RDF files were produced for each thesaurus. The newly created files were first validated using W3C RDF validation service. This is a basic RDF syntax validation test, and all files passed this

initial run with no errors or warnings. The files were then checked using the W3C SKOS validation service [15]. This consists of a series of SKOS compatibility and thesaurus integrity tests, and the output was a set of validation reports. A few minor anomalies arose from these tests, including legacy features such as orphan concepts.

The conversion is efficient and reliable so any updates to thesaurus data at source can be quickly reprocessed. The resultant SKOS files are intended as data inputs to the STAR project and will be used for query expansion and domain navigation tools. It is notable that the validation made possible by the SKOS conversion proved useful to the thesaurus developer for maintenance purposes.

3.3 SKOS based Terminology Services

An initial set of semantic web services have been developed, based upon the SKOS thesaurus representations. These were integrated with the DelosDLMS prototype next-generation Digital Library management system [1]. The services provide term look up, browsing and semantic concept expansion [2]. A pilot SKOS service should shortly be available on a restricted basis from the Glamorgan website. Details of the API and a pilot demonstrator can be found off the STAR website under Semantic Terminology Services [17].

The service is written in C#, running on Microsoft .NET framework and is based on a subset of the SWAD Europe SKOS API, with extensions for concept expansion. The services currently provide term look up across the thesauri held in the system, along with browsing and semantic concept expansion within a chosen thesaurus. This allows search to be augmented by SKOS-based vocabulary and semantic resources (assuming the services are used in conjunction with a search system). Queries may be expanded by synonyms or by semantically related concepts. For example, a query is often expressed at a different level of generalisation from document content or metadata, or a query may employ semantically related concepts. Semantic expansion of concepts for purposes of query expansion yields a ranked list of semantically close concepts [19].

4 Mapping between SKOS and other representations

The next phase of the STAR project involves connecting the thesauri expressed in SKOS to documents or data base items and to an upper ontology, the CIDOC CRM. Figure 2 shows the current model for integrating the thesauri with the CRM. This illustrates two issues concerning the exploitation of SKOS RDF data: (a) the connection between a SKOS concept and the data item it represents and (b) the connection between the CRM and SKOS.

(a) Connecting SKOS concepts and data

The connection between a SKOS concept and an information item is here modeled by a project specific *is represented by* relationship (Figure 2). This is chosen as being the most flexible possibility, which can, if needed, be modified to take account of any standards developments in this area. Another possibility might be the standard *DC: Subject of* if that were appropriate. However, in STAR the application to data items is arguably not quite the same relationship. Another issue is whether, and to what extent, this *concept-referent* relationship should be modeled in SKOS, as opposed to some other indexing or vocabulary use standard. In addition to distinguishing between indexing and classification use cases, there are various other novel DL use cases where KOS are applied to non-traditional data sets for non-traditional purposes. It is important to note the difference between Library Science KOS (intended for information retrieval purposes) and many AI ontology applications, which aim to model a mini-world, where the connection is commonly taken to be a form of *Instance* relationship [18].
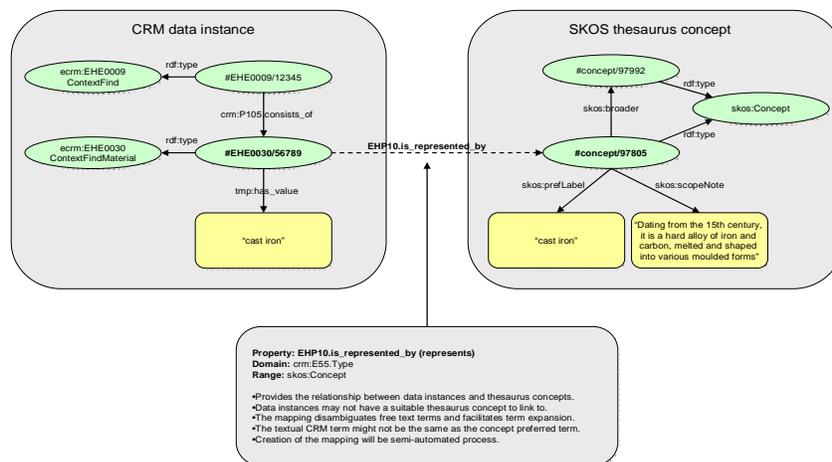

Figure 2: Model for combining SKOS and CIDOC CRM

(b) Connecting SKOS concepts and an upper ontology

The appropriate connection between an upper ontology and domain thesauri or other information retrieval KOS depends upon the intended purpose. It also depends on the alignment of the ontology and domain KOS, the number of different KOS intended to be modeled and the use cases to be supported. Cost benefit issues are highly relevant. This is similar to the considerations and likely success factors for mapping between thesauri or KOS generally (for more details, see the discussion in [13, Section 6.2.1].

In some situations, where the aim is to support automatic inferencing, it may be appropriate to formalize the domain KOS and completely integrate them into a formal ontology, expressing the KOS in OWL, for example. This would allow any benefits of inferencing to be applied to the more specific concepts of the domain KOS. This, however, is likely to be a resource intensive exercise. Since information retrieval KOS and AI ontologies tend to be designed for different purposes, this conversion may change the underlying structure and the rationale should be considered carefully. The conversion may involve facet analysis to distinguish orthogonal facets in the domain KOS, which should be separated to form distinct hierarchical facets. It may involve modeling to much more specific granularity of concepts if the upper ontology is intended to encompass many distinct domain KOS; for example, the need for disambiguation may well not be present in the KOS considered separately but is required when many are integrated together.

Such highly specific modeling should be considered in terms of costs and benefits. It is important to consider the use cases driving full formalisation, since information retrieval KOS, by design, tend to express a level of generality appropriate for search and indexing purposes and driving down to greater specificity may yield little cost benefit for retrieval or annotation use cases. It can be argued that SKOS representation offers a cost effective approach for many annotation, search and browsing oriented applications that don't require first order logic. The SWDWG is currently discussing the recommended best practice for combining SKOS and OWL, following the principle of allowing as many different application perspectives and use cases, as is consistent with the respective underlying principles.

A variant of the above approach, which allows the easier option of SKOS representation, is to consider the domain KOS as leaf nodes of an upper ontology, expressing this, with some form of *subclass* or *type* relationship, depending on the degree of confidence in the mapping. This corresponds to *Leaf Node Linking* in Zeng & Chan's review of mapping [20]. In the CIDOC CRM, for example, one recommended approach is to assert an Instance relationship between a Type property of a CRM class and the top of a thesaurus hierarchy (or the top concept of an entire KOS).

In some cases, including (initial analysis of) the EH case study described above, the domain thesauri may not fit neatly under the upper ontology, the thesauri being designed separately for different purposes. In the STAR project, from the initial discussions with EH collaborators with a subset of the thesauri, the appropriate connection may be a looser SKOS *mapping (broader)* relationship between groups of concepts rather than complete hierarchies. Yet another possibility can be found in Figure 2, which shows a data instance mapped to a CRM entity and where the data items are also indexed with thesaurus concepts. In this case, there is a mapping between data and the integrating upper ontology and another mapping between database fields and the domain thesaurus.

The appropriate mapping between domain thesaurus and the upper ontology ultimately rests upon the use cases to be supported by any explicit connection. In

general, these would tend to be use cases based upon either interactive browsing or automatic expansion (reasoning) of the unified concept space.

Conclusions

The modular approach (coupled with the uniform ID format used) facilitated extraction and storage of relational data into separate RDF files based on CRM-EH structure, and allowed the subsequent merging of selected parts of the data structure originating from multiple data sets. Further combining this data with the CRM-EH ontology (itself a modular unit extending the existing CIDOC CRM) opens up the possibility of automated traversal across known relationships. While more work needs to be done investigating scalability and performance issues, this illustrates potential as a foundation data structure for a rich application.

A CRM based web service has been implemented over the extracted data and model, which offers search capability with subsequent browsing over CRM-EH relationships.

Acknowledgements

References

1. Binding C., Brettlecker G., Catarci T., Christodoulakis S., Crecelius T., Gioldasis N., Jetter H-C., Kacimi M., Milano D., Ranaldi P., Reiterer H., Santucci G., Schek H-G., Schuldt H., Tudhope D., Weikum G.: DelosDLMS: Infrastructure and Services for Future Digital Library Systems, 2nd DELOS Conference, Pisa (2007) http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334
2. Binding C., Tudhope D.: KOS at your Service: Programmatic Access to Knowledge Organisation Systems. Journal of Digital Information, 4(4), (2004) http://journals.tdl.org/jodi/article/view/jodi-124/109
3. CIDOC Conceptual Reference Model (CRM), http://cidoc.ics.forth.gr
4. CRM-EH Extension to CRM http://hypermedia.research.glam.ac.uk/kos/CRM/
5. 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
6. DELOS Cluster on Knowledge Extraction and Semantic Interoperability, http://delos-wp5.ukoln.ac.uk/
7. Doerr, M.: The CIDOC Conceptual Reference Module: an Ontological Approach to Semantic Interoperability of Metadata. AI Magazine, 2493, 75--92 (2003)
8. English Heritage http://www.english-heritage.org.uk/

9.   English Heritage Thesauri http://thesaurus.english-heritage.org.uk/
10. May, K.: Integrating Cultural and Scientific Heritage: Archaeological Ontological Modelling for the Field and the Lab. CIDOC CRM Sig Workshop, Heraklion (2006) http://cidoc.ics.forth.gr/workshops/heraklion_october_2006/may.pdf
11. Nußbaumer, P., Haslhofer, B.: CIDOC CRM in Action – Experiences and Challenges. Poster at 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL07), Budapest (2007) http://www.cs.univie.ac.at/upload//550/papers/cidoc_crm_poster_ecdl2007.pdf
12. Nußbaumer, P., Haslhofer, B.: Putting the CIDOC CRM into Practice – Experiences and Challenges. Technical Report, University of Vienna (2007) http://www.cs.univie.ac.at/publication.php?pid=2965
13. 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)
14. SEMWEB RDF Library for .NET, http://razor.occams.info/code/semweb
15. SKOS: Simple Knowledge Organization Systems, http://www.w3.org/2004/02/skos
16. SKOS API. SWAD_EUROPE Thesaurus Project Output (2004) http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html
17. STAR Project: Semantic Technologies for Archaeological Resources, http://hypermedia.research.glam.ac.uk/kos/star
18. Tudhope D., Koch T., Heery R.: Terminology Services and Technology: JISC State of the art review (2006) http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_te chnology_review_sep_06.pdf
19. Tudhope D., Binding C., Blocks D., Cunliffe D. Query expansion via conceptual distance in thesaurus indexed collections. Journal of Documentation, 62 (4), 509–533. Emerald (2006)
20. Zeng M, Chan L. Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5): 377 – 395. (2004)