Project no.507618

DELOS

A Network of Excellence on Digital Libraries

Instrument: Network of Excellence

Thematic Priority: IST-2002-2.3.1.12

Technology-enhanced Learning and Access to Cultural Heritage

"Making KOS Machine Understandable"

Additional Report for Work Package 5

JPA4 Period January – December 2007

Submission date: 29 February 2008

Start Date of Project: 01 January 2004

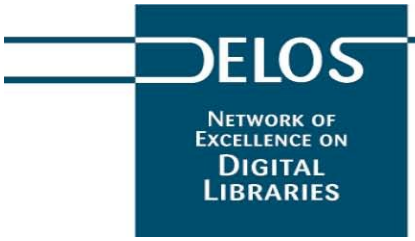
Duration: 48 Months

UKOLN, University of Bath

Draft

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Dissemination Level: [PU (Public)



Machine Understandable

Knowledge Organization Systems[1]

Authors: Doug Tudhope, Ceri Binding, University of Glamorgan

[1] Report prepared for DELOS WP5 in response to *7) Proposal of method for the mapping of knowledge organisation systems (sets of controlled vocabulary migrated to RDF/OWL and /or SKOS according to short term actions as recommended by the i2010 Interoperability Group) as part of WP5*

Document status

Date	Version	Modifications
February	**Draft**	Circulated to selected Cluster partners.
February	**Final Draft**	Circulated to Scientific Board

TABLE OF CONTENTS

(8) Basic Semantic Interoperability

Make existing metadata and the controlled terminology used therein machine understandable to create a data layer ready for semantic query methods. The method of choice for conversion is SKOS, but use of OWL or RDF may be appropriate in some application scenarios.

S. Gradman, i2010 Interoperability Group: Short term Agenda Issue[2]

1. Introduction

Knowledge Organization Systems (KOS) are part of an attempt improve access to digital resources via vocabulary control and knowledge organization. Vocabulary control aims to reduce the ambiguity of natural language when describing and retrieving items for purposes of information searching. When searching free text with uncontrolled terms, significant differences can stem from trivial variations in search statements and from differing conceptualisations of an information need. Different people use different words for the same concept or employ different concepts. Controlled vocabularies consist of terms, words from natural language selected as useful for retrieval purposes by the vocabulary designers. A term can be one or more words and is taken to represent a concept.

Two features of natural language pose significant problems for information systems; different terms can represent the same concept, while the same term can ambiguously represent different concepts. Controlled vocabularies attempt to reduce ambiguity by defining the scope of terms and more complex vocabularies provide a set of (effective) synonyms for each concept. They may also provide vocabulary for KOS, which organise and structure their concepts via different types of semantic relationship. The presentation of concepts in hierarchies and other semantic structures helps both indexer and searcher choose the most appropriate concept for their purposes. This semantic structure also affords a mechanism (for both human and machine) to connect a searcher with an indexer's choice of terminology by traversing the pathways.

There are many different kinds of KOS, with different degrees of vocabulary control, richness of semantic relationships and formality - designed to serve different purposes. These include classification systems, gazetteers, lexical databases, ontologies, taxonomies and thesauri. This report focuses upon the family of information retrieval KOS, with semantics designed to support annotation, browsing, indexing and search use cases. For a review of different kinds of KOS, see Tudhope *et al.* (2006).

[2] From Gradman S. 2007, Presentation on Interoperability of Digital Libraries. Report on the work of the EC working group on DL interoperability.

SKOS Core is a W3C Working Draft RDF/XML representation for KOS, developed and maintained by the W3C Semantic Web Deployment Working Group (SWDWG). It was originally conceived with thesauri in mind but the intention is to encompass other structured KOS, such as taxonomies and classifications, and less structured vocabularies for social tagging and Web 2.0 applications. The SKOS website has links to a Quick Guide to Publishing a Thesaurus on the Semantic Web and to a SKOS Reference Working Draft (SKOS). There is also a list of some of the SKOS representations of KOS to date (SKOS Data Zone).

This report focuses upon issues concerning the production of machine readable SKOS representations of thesauri (and other information retrieval KOS), SKOS being the i2010 group's preferred method of conversion. Basic methods of SKOS conversion are outlined and illustrated by a case study from WP5 activities. Important associated issues are briefly mentioned, including the 'mapping' between SKOS concepts and information (data) items and between SKOS concepts and an upper ontology. The provision of standard SKOS representations makes possible the provision of common SKOS-based web services and this is briefly noted.

1.1. Simple Knowledge Organization System (SKOS)

SKOS is an RDF/XML representation standard based on a formal data model. It is intended for the large family of vocabularies and concept structures, with a lightweight semantics designed for information retrieval purposes, rather than formal logic (Miles *et al.* 2005). Reports of some early work leading up to the development of SKOS include Cross *et al.* (2000), Miles *et al.* 2004, Wilson & Mathews (2002).

For reports of experience in extending KOS to AI ontology (OWL) models, see for example (Soergel *et al.* 2004, van Assem *et al.* 2004, Wielinga *et al.* 2001). It should be noted that SKOS is intended to be compatible with OWL. The SWDWG is currently considering the appropriate mechanisms for combining SKOS and OWL representations, where for example a formal ontology is combined with SKOS representations of information retrieval KOS.

2. Conversion of KOS to SKOS/RDF representations

In the long run it is possible that vendors of editors for thesaurus and other KOS will include SKOS as an output format, while major vocabulary providers will include a SKOS version in distributions. Currently this is not the case and it will be necessary to convert legacy KOS from other formats to SKOS for semantic interoperability purposes.

There is some considerable variation in the features supported by different thesauri and some thesauri do not adhere to the thesaurus standards, employing for example specific, nom-standard relationships or properties. There is also a distinction between (older) term-based thesauri and concept based thesauri that follow the more recent standard versions. Van Assem *et al.* (2006) outline a general method of converting thesauri to SKOS and report on experiences with three quite different thesauri. The first step in their method

entails an analysis of the thesaurus to determine whether there are any non-standard features. In some cases, non-standard features can be accommodated by specialization of the core SKOS elements. Other departures from the standards may entail some loss of original features in the SKOS version; they note examples of concept-term relationships in MeSH which proved difficult to model in the version of SKOS employed. They note the need to provide unique identifiers for the SKOS representation and that URIs may need to be invented if the thesaurus has no notion of identifiers. They also note the need in some cases for consultation with the thesaurus providers or experts on the intended aim of non-standard features. After analysis of thesaurus elements and decision on a strategy for non-standard features, different conversion routes were employed for different thesaurus formats: an SWI-Prolog program converted from XML format to SKOS RDF, a Perl program from plain text,

If the KOS is available in an XML representation, conforming to a published XML Schema, then conversion is facilitated and it may even be possible to derive an XSL transform to achieve the conversion. The latter option assumes that a thesaurus, for example, conforms closely with the thesaurus standards and is reasonably compatible with the SKOS data model. If vocabulary representations are based on an underlying formal model then it is easier to derive transformations to particular syntactical representations, although significant analysis and judgments may be needed where the models are incompatible. Relevant XML Schema include the XML Schema for the new BSI Thesaurus Standard (8723), the MARC 21 Format for Authority Data, the Zthes 1.0 XML Schema and various vendor XML formats. In addition to the XML Schema, the BS8723-5 (interoperability) working group development website (BS8723-5) holds examples of thesaurus representations and various XSL conversions from one Schema to another. A conversion to SKOS is listed as a possible future task. Part 5 of the new BSI Guide on Exchange formats and protocols for interoperability will be published shortly.

While this is an attractive option, there are various issues which may prove problematic in some situations with this approach. It may sometimes be important to take into account character encodings. For example, Vizine-Goetz *et al.* (2006) discuss problems automatically converting between MARC-XML and SKOS RDF-XML, where it was found necessary to employ an XSLT 2.0 processor to create an XSL 2.0 transform due to differences in character encodings. Concept or term identifiers may also pose problems since some vocabularies may lack unique IDs or may not have Web actionable URLs. Another issue is that many XML parsing applications employ the XML DOM and require the entire XML document to be stored in memory. This causes problems when an XML distribution is too large for available memory resources. In such situations, it may be possible to employ an XSL transform, using tools such as Saxon (v9) to select only those elements of an XML distribution required.

Another approach is to import the XML distribution into a database and create a custom SKOS output generator from the database. In some cases, the KOS will be available in a relational database format. This approach may also be necessary in situations where the KOS is distributed as a spreadsheet or CSV file, etc. and it is easiest to import it into a relational database. Once a standard database schema has been established then SKOS can be produced for any KOS that has been imported into that structure.

The following case study illustrates recent (2007) experience by the (University of Glamorgan) STAR project in producing SKOS representations for several English Heritage (EH) thesauri. It illustrates many of the issues discussed above, including an initial XSL transform approach that worked well for smaller thesauri but not the larger ones. It also illustrates issues where a KOS may not map completely to the SKOS model. In general, thesauri conforming to the BSI/NISO/ISO standards should map in a fairly straight forward manner to SKOS. As noted above, 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 case study concern the need to create URIs for concept identifiers as part of the conversion and the potential for validation.

As mentioned above, SKOS was originally developed for thesauri but the scope has been widened to orient to other structured KOS, such as taxonomies and classifications, and also less structured vocabularies for social tagging and Web applications. SKOS allows for specialization and extension of the core model. One issue will be whether other types of KOS require specialization of SKOS. For example, some simple taxonomies may be encompassed within SKOS with relatively little specialization, if any. However, complex classification schemes will require specializations and extensions if their full content is to be captured.

3. STAR Project Case Study

For the purposes of the STAR project it was required to initially convert any received thesaurus data into a common standard format. The output format chosen for this exercise was SKOS RDF.

3.1 Data received

Thesaurus data was received from English Heritage National Monuments Record Centre, in 5 CSV format files. The file structures represented the 5 tables in the Thesaurus Management module of the English Heritage AMIE database from where the data had been exported, with the first row in each file holding the column names:

- classification_groups.csv
- thesaurus_terms.csv
- thesaurus_term_relation.csv
- thesaurus_term_preferences.csv
- thesaurus_term_uses.csv

3.2 Conversion process

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 (EH2SKOS.exe) was written to export the data from this database into SKOS

RDF format. The overall procedure followed is illustrated in Figure 1 (which also encompasses some other STAR project activities).

The main caveat 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.

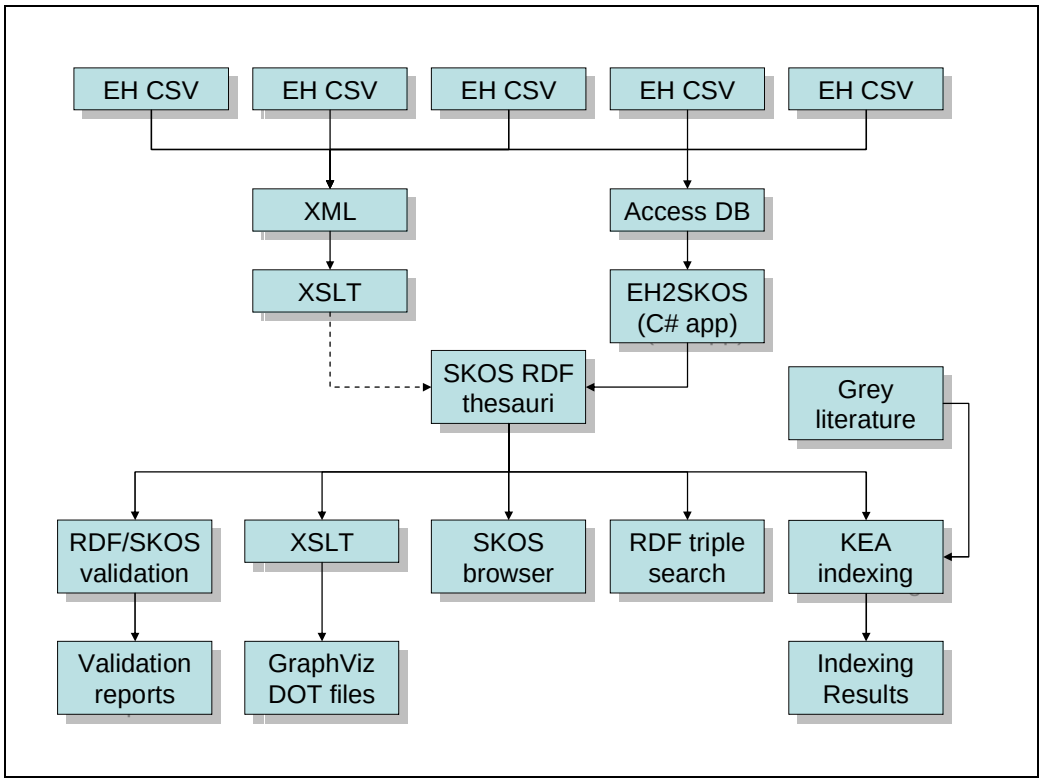


Figure 1 – EH CSV to SKOS conversion and validation process

3.3 Validation process

As a result of running the conversion application, 6 separate RDF files were produced, one 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 (see SKOS Website). This consists of a series of SKOS compatibility and thesaurus integrity tests, and the output was a set of validation reports. A few

anomalies arose from these tests requiring further investigation by the EH thesaurus developers, 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 KOS developer for these maintenance purposes.

A small C# application was written based on the Open Source "Drive" RDF parser. This application was able to import the SKOS files and traverse the RDF graph structures, producing a summary count of the various relationship types contained within each RDF graph, as shown in Figure 2.

File name	**Thesaurus Description**	**Concepts**	**Top concepts**	**BT relationships**	**NT relationships**	**RT relationships**	**ALT/UF terms**	**Scope notes**
EH1_20070130.rdf	Monument types	3987	18	5281	5281	8050	2837	3949
EH92_20070130.rdf	Evidence thesaurus	41	1	40	40	8	31	41
EH128_20070130.rdf	MDA Object types	1742	24	1916	1916	222	509	1726
EH129_20070130.rdf	Main building materials	225	9	221	221	242	438	225
EH131_20070130.rdf	Covering building materials	225	9	220	220	246	437	224
EH560_20070215.rdf	Archaeological sciences	101	5	98	98	6	21	96
Totals		**6321**	**66**	**7776**	**7776**	**8774**	**4237**	**6261**

Figure 2 – Statistics for SKOS thesauri produced by conversion process

3.4 SKOS based Terminology Services

The STAR project has developed an initial set of semantic web services, based upon the SKOS thesaurus representations. These were integrated with the DelosDLMS prototype next-generation Digital Library management system, built on the OSIRIS middleware environment (ETH Zurich and University of Basel) and an account was published in the Second DELOS Conference in Pisa Proceedings (Binding et al. 2007). The services provide term look up, browsing and semantic concept expansion (Binding & Tudhope 2004). A pilot SKOS service should shortly be available on a restricted basis from the Glamorgan website (http://hypermedia.research.glam.ac.uk/kos/terminology_services).

The service is written in C#, running on Microsoft .NET framework (version v2.0.50727) and is based on a subset of the SWAD Europe SKOS API, with extensions for concept expansion. The service consists of 8 function calls, which can be integrated into a textual

or metadata based search system. Functionality includes a facility to look up a user provided string in the controlled vocabularies of all KOS known to the server, returning all possibly matching concepts. The ability to browse concepts via the semantic relationships in a thesaurus is provided. Semantic expansion of concepts for purposes of query expansion is also possible; (configurable) automatic traversal of SKOS relationships yields a ranked list of semantically close concepts.

The services allow search to be augmented by KOS-based vocabulary and semantic resources. Users may browse a concept space to explore and become familiar with specialist terminology or may browse to directly access data linked to concepts. 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. This provides an augmented textual search capability to complement existing OSIRIS content-based retrieval.

3.5 Mapping between SKOS and other representations

Part 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 3 shows the current model for integrating the thesauri with the CRM, which has been extended by EH to model the excavation and analysis workflow. 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 3). 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 information retrieval KOS 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.

(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 Patel e*t al.* 2005, 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 is not present in the KOS considered separately but is required when they are integrated together.

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 annotation, search and browsing oriented applications that don't require first order logic (Tudhope & Binding 2008). 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 (2004). In the CIDOC CRM, for example, one recommended approach is 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 so neatly with the upper ontology, the thesauri being designed separately for different purposes. 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. However, an alternative connection is represented in Figure 3. This shows a data instance mapped to a CRM entity, where data items are also indexed with thesaurus concepts (or the database glossaries or pick lists can be mapped to 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 case for an additional mapping between domain thesaurus and the upper ontology rests upon the precise use cases to be supported by the 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.

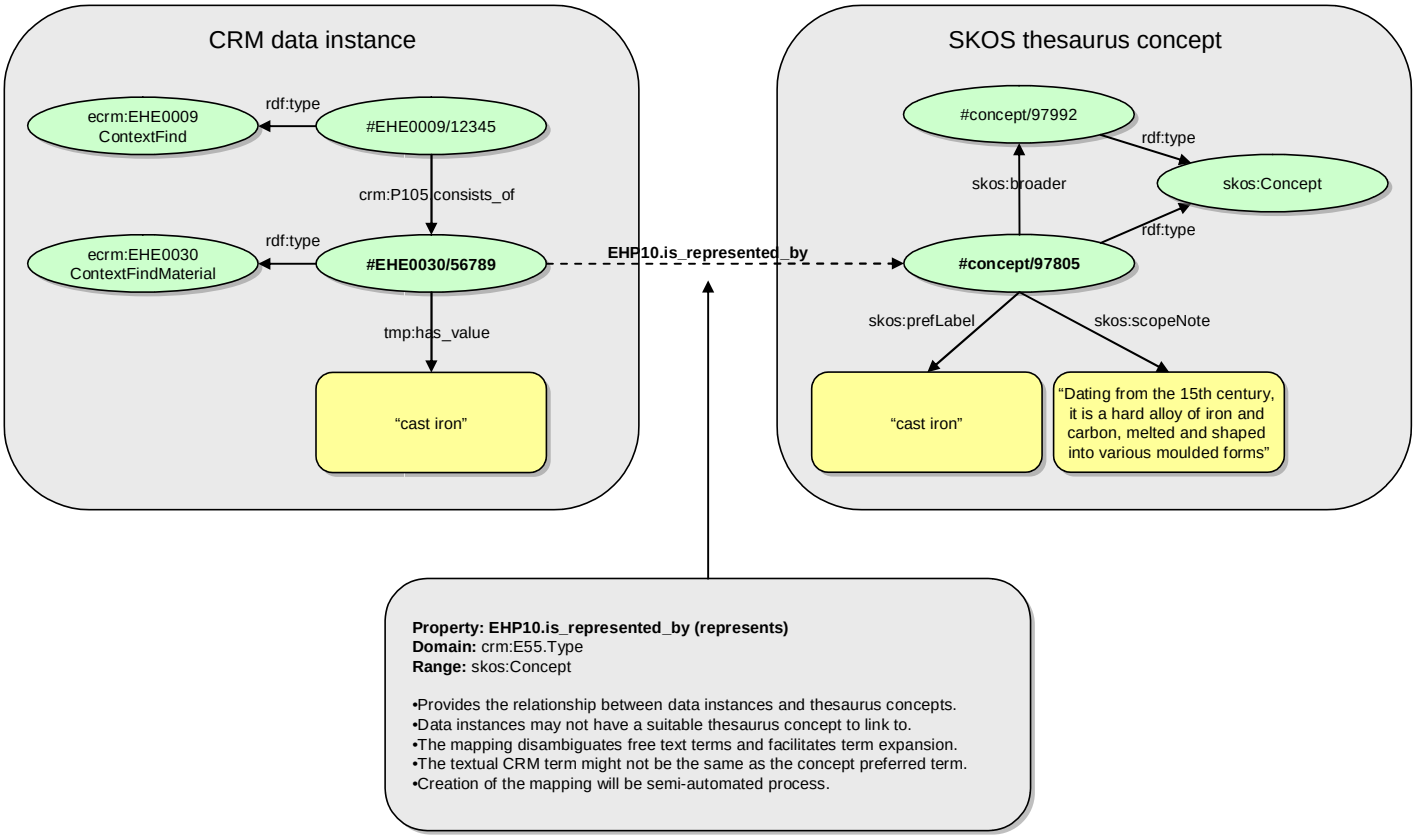


Figure 3. Connections between CRM data instance and SKOS concept

Acknowledgements

The STAR project is funded by the UK Arts and Humanities Research Council (AHRC). Thanks are due to Andrew Houghton (OCLC Research) for helpful input to various parts of the report and to Phil Carlisle, Keith May (English Heritage).

References

AMIE - Archives and Monuments Information England (http://www.english-heritage.org.uk/upload/pdf/A_Quick_Guide_to_AMIE.pdf)

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. 2007. DelosDLMS: Infrastructure and Services for Future Digital Library Systems, *Proceedings 2nd DELOS Conference*, Pisa. http://www.delos.info/index.php?option=com_content&task=view&id=602&Itemid=334

Binding C., Tudhope D. 2004. KOS at your Service: Programmatic Access to Knowledge Organisation Systems. *Journal of Digital Information,* 4(4), http://journals.tdl.org/jodi/article/view/jodi-124/109

BSI 8723. Structured vocabularies for information retrieval — Guide — British Standards Institution. - London : BSI, 2005.

BS8723-5 Thesaurus Standard Development Website. http://schemas.bs8723.org/

Cross P., Brickley D., Koch. T. 2000. Conceptual relationships for encoding thesauri, classification systems and organised metadata collections and a proposal for encoding a core set of thesaurus relationships using an RDF Schema. *DESIRE Project Report.* http://www.desire.org/results/discovery/rdfthesschema.html

English Heritage http://www.english-heritage.org.uk/

English Heritage Thesauri http://thesaurus.english-heritage.org.uk/

MARC 21 XML Schema http://www.loc.gov/standards/marcxml/

Miles A., Matthews B., Wilson M. 2004. RDF encoding of multilingual thesauri. *Deliverable 8.3*, version 0.1, SWAD-Europe Project. http://www.w3c.rl.ac.uk/SWAD/deliverables/8.3.html.

Miles A., Mathews B., Wilson M. 2005. SKOS Core: Simple Knowledge Organisation for the Web, Alistair Miles, Brian Matthews and Michael Wilson, *Proceedings of the International Conference on Dublin Core and Metadata Applications,* (DC 2005), 5-13

Patel M., Koch T., Doerr M., Tsinaraki C. 2005. *Report on Semantic Interoperability in Digital Library Systems.* DELOS Network of Excellence, WP5 Deliverable D5.3.1.

Soergel D., Lauser B., Liang A., Fisseha F., Keizer J., Katz S. 2004. *Journal of Digital Information* 4(4), Reengineering Thesauri for New Applications: the AGROVOC Example. http://journals.tdl.org/jodi/article/view/jodi-126/111

SAXON. http://saxon.sourceforge.net/

STAR Project - Semantic Technologies for Archaeological Resources http://hypermedia.research.glam.ac.uk/kos/star

SKOS - Simple Knowledge Organization Systems http://www.w3.org/2004/02/skos

SKOS Data Zone - http://esw.w3.org/topic/SkosDev/DataZone

SKOS API. 2004. SWAD_EUROPE Thesaurus Project Output. http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html

SWDG - Semantic Web Deployment Working Group. http://www.w3.org/2006/07/SWD

Tudhope D., Koch T., Heery R. 2006. Terminology Services and Technology: *JISC State of the art review.* http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

Tudhope D., Binding C. 2008 forthcoming. Faceted Thesauri. *Axiomathes*, Springer. http://dx.doi.org/10.1007/s10516-008-9031-6

van Assem A., Malaisé V., Miles A., Schreiber G. 2006. A Method to Convert Thesauri to SKOS. *Proceedings European Semantic Web Conference,* Springer LNCS 4011, 95-109

van Assem A., Menken M., Schreiber G., Wielemaker J., Wielinga B. 2004. A Method for Converting Thesauri to RDF/OWL. *Proceedings International Semantic Web Conference,* Springer LNCS 3298, 17-31

Vizine-Goetz D., Houghton A., Childress E., "Web Services for Controlled Vocabularies", *ASIS&T Bulletin,* June/July 2006, Available online at http://www.asist.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html

W3C RDF validation service (http://www.w3.org/RDF/Validator)
W3C SKOS validation service (http://www.w3.org/2004/02/skos/validation)

Wielinga B., Schrieber A., Wielemaker J., Sandberg J. 2001. From Thesaurus to Ontology. *Proceedings 1st International Conference on Knowledge Capture (K-CAP'01).* 194-201. ACM Press

Wilson M., Matthews B. 2002. Migrating Thesauri to the Semantic Web. Article. *ERCIM News* No. 51, http://www.ercim.org/publication/Ercim_News/enw51/wilson.html

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), pp. 377 – 395, 2004.

Zthes. http://zthes.z3950.org/