Towards KOS-based semantic services

Doug Tudhope
Hypermedia Research Unit
University of Glamorgan

Helsinki, November 30, 2007

Presentation

- Possible approaches to interoperability (from Nov 29)
  - Standards
  - Combination of KOS
- STAR – combining KOS and core Ontology
  Semantic Technologies for Archaeological Resources
- EnTag – combining KOS and social tagging 'folksonomies'
- SKOS (standards) based services

STAR
Semantic Technologies for Archaeological Resources

ENGLISH HERITAGE

Arts & Humanities Research Council

DANMARKS BIBLIOTEKSSKOLE

Acknowledgement Ceri Binding for some of the STAR material

Project Outline

§ 3 year AHRC funded project
§ Started January 2007, finish December 2010
§ Collaborators
  § English Heritage
  § RSLIS Denmark
§ Aim – "*To investigate the potential of semantic terminology tools for widening access to digital archaeology resources, including disparate datasets and associated grey literature*"

Background

- Current EH situation one of fragmented datasets and applications, with different terminology systems
- Interpretation may not consist of same terms as context
- Searchers from different scientific perspectives may not use same terminology
- Need for integrative metadata framework
  EH have designed an upper ontology based on CRM standard
- Work to date focused on modelling

Databases not meaningfully connected

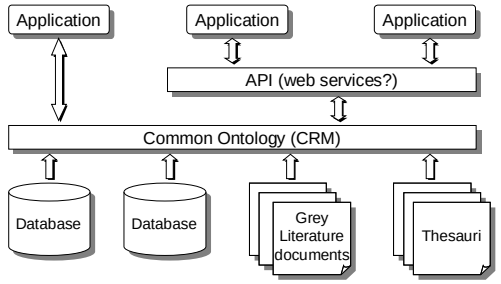
- Even simply expressed queries currently difficult to answer, due to lack of tools for cross database searching

"Specialists could only talk to [field] archaeologists and not talk to each other".

(from discussion with a palaeoenvironmental archaeologist)

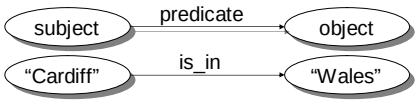
Wider questions arising from science analysis by finds specialists often referred back to field archaeologist since databases documenting different scientific aspects not meaningfully connected

General architecture



Resource Description Framework (RDF)

§ http://www.w3.org/RDF/
§ XML / URI based format
§ Modelling of graph structures
§ RDF triples:



Simple Knowledge Organisation Systems (SKOS)

§ STAR thesauri represented in SKOS

§ http://www.w3.org/2004/02/skos/
§ *Formal* representation of thesauri, taxonomies, classification schemes etc. in RDF/XML with *looser* semantics than OWL

cost/benefit advantages for SKOS for STAR purposes
§ Semantics at a suitable level for retrieval purposes
§ Less overhead to set up
§ Can apply concept-based semantic expansion from FACET (see Nov 29)

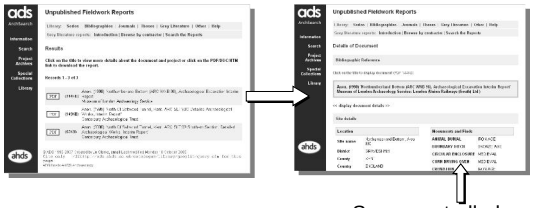
CIDOC Conceptual Reference Model (CRM)

§ "A reference ontology for the interchange of cultural heritage information" [http://cidoc.ics.forth.gr/]
§ International standard ISO 21127:2006

§ Extensions by English Heritage
Modelling workflow of excavation process and analysis

§ CRM as overarching framework
§ We use RDFS representation

ADS grey literature online

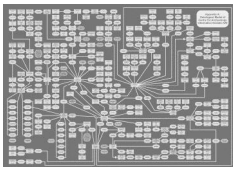


•http://ads.ahds.ac.uk/catalogue/

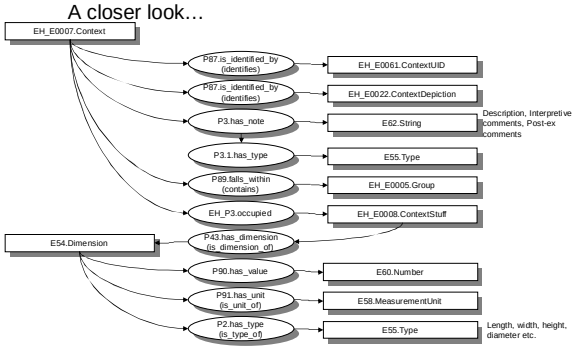
Some controlled vocabulary indexing

EH extension to CRM

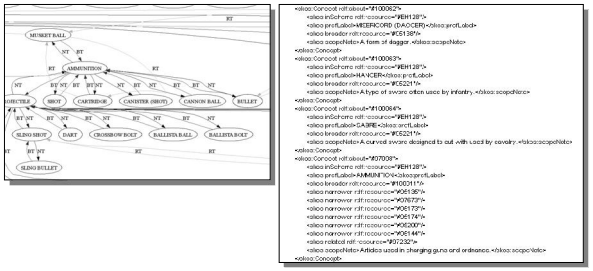
- Currently in pdf file
- Need to represent in machine readable format



A closer look…



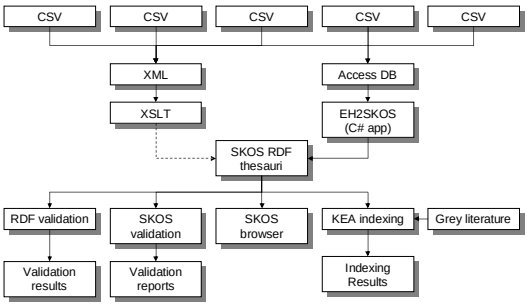
Representing Thesauri in SKOS RDF



English Heritage Thesauri

- Monument types thesaurus
  - Classification of monument type records
- Evidence thesaurus
  - Archaeological evidence
- MDA object types thesaurus
  - Archaeological objects
- Building materials thesaurus
  - Construction materials
- Archaeological sciences thesaurus
  - Sampling and processing methods and materials

English Heritage Thesauri – conversion to SKOS RDF format



Environmental Archaeology Thesaurus
Scope Notes Extract (i)

- ***Altered by Animals***
- SN: Modification or damage by an animal
- RT Worked (use where modification is by humans *in ASPECT*)
- ***Anoxic***
- SN: Material preserved by exclusion of oxygen usually due to saturation with water which inhibits decay by micro-organisms
- Non Preferred Term: Waterlogged
- ***Burnt***
- SN: Use for material that has been burnt
- ***Calcined***
- SN: Material burnt at a high temperature (above 700 degrees centigrade) leaving only the mineral component.
- Non-preferred term: cremated
- BT: Burnt
- RT Cremation
- ***Charred***
- SN: Material that has been burnt and at least in part reduced to carbon as a result of burning in a reducing atmosphere below 500 degrees C.
- Non-preferred term: Carbonised
- BT: Burnt
- ***Silicified***
- SN Use for material that has been burnt at high temperatures in a good air supply such that only silica component remains
- BT: Burnt
- ……
- ***Mineral Replaced***
- SN: Replacement of organic material by minerals, including calcium carbonate and calcium phosphate
- Non Preferred Term: Mineralised, Fossilised
- ***Mineral Preserved***
- SN: Preservation of material by the toxic effect of corrosion products in the immediate vicinity, or within

Example of CRM - Thesauri connection

- FlotationSampleResidueType – EH_E0067
  CRM entity E55: Type

  Classification of flot and/or residue contents
- Mapping (by EH collaborators):
  Use Arch Science Thesaurus Terms:
  Object type, Material type, Modification state, Aspect

Example CRM - Thesaurus connection 2

- ContextSampleType – EHE0053
- CRM entity E55: Type
- Derived from the Environmental guidelines list
  Samples taken will be of a particular type depending upon the technique that will be used to analyse them.
- For Specialist Scientific Sampling it would be appropriate to use Archaeological Science Thesaurus terms for “Investigative Techniques”, but for samples taken by non-specialists the investigative technique may not be known at the point of sampling.

CRM - Thesaurus connection ?

- How to model connection between CRM-EH ontology and the STAR thesauri?
- Initial analysis suggests does not necessarily fit neatly under CRM-EH entities as a direct relationship
- Currently considering applying BSI Part 4 mapping relationships (subset SKOS mapping relationships)

Data modelling

- The data extraction process involved selected data from the following archaeological datasets:
  Raunds Roman Analytical Database (RRAD)
  Raunds Prehistoric Database (RPRE)
  York Archaeological Trust Integrated Archaeological Database (IADB)
- Approach 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 allows data instances to be later selectively combined as required.

Example – data mapping to CRM-EH

CRM-EH	RRAD	RPRE	IADB
EHE0007.Context			
EHE0061.ContextUID	context.ContextNo	contexts.01:RECORD NUMBER	contexts.CONTEXT
EHE0022.ContextDepiction	context.Easting; context.Northing	contexts.04:CO-ORDINATES	contexts.GRIDREF_E; contexts.GRIDREF_N
EHE0046.ContextNote	context.Description; context.InterpretiveComments	contexts.15:DESCRIPTIVE TEXT; contexts.60:INTERPRETATIVE COMMENTS	contexts.DESCRIP; contexts.NOTES
EHE0009.ContextFind			
EHE1004.ContextFindDepositionEvent	Object.WithinContext; Ceramics.context	objects.37:WITHIN; flint.37:WITHIN	clfr.CONTEXT
EHE0043.ContextFindUID	Object.ObjectNo; Ceramics.CeramicsNo;	objects.01:RECORD NUMBER; flint.01:RECORD NUMBER	clfr.ID
EHE0050.ContextFindMaterial	Object_Material.Material; "ceramic"(?)	objects.24:MATERIAL; flint.24:MATERIAL	clfr.MATERIAL
EHE0096.ContextFindGeneralUse	Object.GenericName;Ceramics_Type	objects.05:CATEGORY; flint.05:CATEGORY	clfr.NAME
EHE0097.ContextFindIntendedUse	Object.SimpleName;Ceramics_Type	objects.05:CATEGORY; flint.05:CATEGORY	clfr.NAME
E62.String (As ContextFindNote)	Object.DescriptiveText; Ceramics.Description; Ceramic.InterpretiveComments	objects.16:DESCRIPTIVE TEXT; objects.60:INTERPRETATIVE COMMENTS	clfr.NOTES; clfr.DESCRIP
EHE1001.ContextEvent (for stratigraphic relationships)	Context_Stratigraphy.RelatedContext, Relationship	contexts.BELOW, FILLED_BY, CUT_BY, BUTTED_BY, WITHIN, CONTAINS, BONDED_WITH, SAME_AS, ABOVE, FILL_OF, CUTS, BUTTS	Strat.CON1; CON2

Figure 4 – mapping of data fields to extended CRM

Example – data modelling

These relationships between entities were extracted and modelled in RDF, for each dataset:

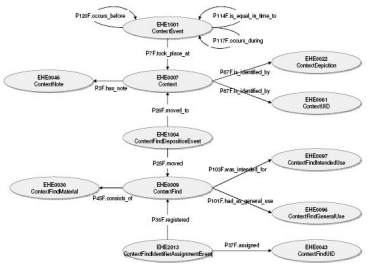


Figure 5 – entity relationships modelled

Potential mapping problems in mapping from different datasets to a general high level ontology

“The first issue is the abstractness of the concepts (e.g. Time Appellation, Man-Made Object) defined by the global ontology, which makes them ambiguous to any human user. Even expert users have produced ambiguous mappings and have required several iterations to produce consistent mapping definitions.

If several experts specify mappings independently from each other, it is very likely that they will produce incompatible mappings and fail the goal of enabling interoperability.

Another point directly connected to the abstractness of the concepts, is the presentation to the user. Basically a graphical user interface is required which hides the complexity of the global ontology and allows the user to formulate queries over more concrete concepts.”

From BRICKS FP6 IP CIDOC CRM Poster at ECDL07, Budapest

Potential mapping problems
in mapping from different datasets to high level ontology
STAR approach

- BRICKS worked with high level basic CIDOC CRM schema
- CRM-EH has detailed *extension* of high level CIDOC CRM entities modelling workflow of excavation and analysis
  Our approach is that this will be easier to map to
- STAR confined to UK archaeology domain
  With mapping done by project team and main collaborators

  Consider generalisation issues as part of project evaluation

Investigate cost/benefit issues …

- Granularity of modelling in CRM-EH ontology
- Granularity of mapping to data
- Extent to which ontology of excavation and analysis process useful for retrieval purposes
- Extent to which ontology can inform user interface
- Scalability of exporting data to RDF
- Generalisability beyond immediate UK context

Presentation

- Possible approaches to interoperability
  - Standards
  - Combination of KOS
- STAR – combining KOS and core Ontology
- EnTag – combining KOS and social tagging 'folksonomies'
- SKOS (standards) based services

EnTag project
Enhanced Tagging for Discovery

- Ongoing JISC funded project to investigate the combination and comparison of controlled and folksonomy approaches to semantic interoperability in the context of repositories and digital collections
- Partners UKOLN, Glamorgan, CCLRC, Intute, with unfunded support from OCLC Research and Royal School of Library and Information Science, Denmark

Glamorgan role
- Develop and evaluate a Demonstrator combining Dewey and social tagging

EnTag project
Enhanced Tagging for Discovery

- Social tagging applications hold promise of reducing indexing costs by drawing end-users into contributing this resource.
- However existing social tagging applications have not been designed with information discovery and retrieval in mind. The resulting folksonomies are completely uncontrolled, lacking even basic control of word forms, spelling, synonyms and disambiguation of homonyms.
- EnTag aims to compare a 'vanilla' social tagging application for Intute Social Sciences collection with an advanced hybrid application, where Dewey terminology resources are employed to structure and 'improve' the indexing, while retaining social tagging nature

Evaluation of EnTag:
considerations of test design

Marianne Lykke Nielsen
Information Interaction and Architecture
Royal School of Library and Information Science

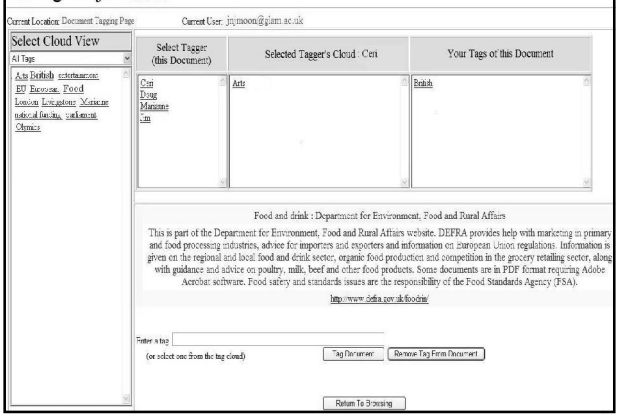
Evaluation – focus and objective

• Context: tagging as part of information searching and relevance assessment, tagging for recommendation and sharing
• Hybrid system: investigate whether tagging can be improved by a combination of traditional tag clouds and clouds of controlled descriptors
- Improve tagging
  - Relevance of tags (perspective, aspects, specificity, exhaustivity, terminology (linguistic level, semantic level, contextual level)
  - Consistency
  - Efficiency (time used)
  - Use (tags selected, clouds consulted, order of consultation)
- Improve retrieval
  - Effectiveness (degree of match between user and system terminology)

Evaluation – test setting

• Comparison test: comparison between 1) control, "vanilla" system and 2) experimental, hybrid system. Open test - taggers know that they use and evaluate two systems
• Number of taggers: 50 taggers, post graduate students using the Intute Social Science as part of common information seeking strategy
• Number of documents: 100 documents, covering up to four topics of relevance for the group of taggers
• Initial Draft Test design:
- Each tagger tags all documents:
  - System 1: 50 documents
  - System 2: 50 documents
- Each document is tagged by all 50 taggers:
  - System 1: 25 taggers
  - System 2: 25 taggers
- 5000 tagging sessions
  - System 1: 2500
  - System 2: 2500

EnTag Project SST



Presentation

- Possible approaches to interoperability
  - Standards
  - Combination of KOS
- STAR – combining KOS and core Ontology
- EnTag – combining KOS and social tagging 'folksonomies'
- SKOS (standards) based services

SKOS API

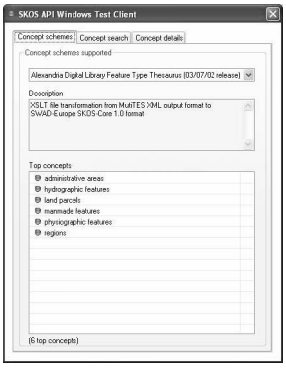
- SKOS API a deliverable of SWAD-Europe Thesaurus Activity - http://www.w3.org/2001/sw/Europe/reports/thes
- SKOS API designed to provide programmatic access to thesauri and related KOS in SKOS Core
- Example SKOS API calls
  - getConcept (uri)
  - getConceptsMatchingKeyword/*Regex* (string)
  - getAllConceptRelatives (concept)
  - getSupportedSemanticRelations
  - getAllConceptRelatives (concept, relation)
  - getAllConceptsByPath (concept, relation, distance)

Pilot KOS Browser Client Web Service

Developed C# client application for SKOS thesaurus server

A 'rich client' browser displays details for SKOS concepts via web service calls

Uses subset of SKOS API with extension for semantic expansion



Also ongoing

- Javascript (AJAX) interface widgets for desktop access to SKOS-based services
- HTML with JavaScript and CSS styling, using Javascript AJAX calls to communicate with the same (SOAP) web service that the C# demonstrator uses
- Contact if interested in trying out early prototypes

Future issues

More complex services as API protocol elements:

- more advanced natural language functionality
- cross-mapping provision
- data-dependent filters (such as number of postings)
- ***semantic expansion as a service***
  - different configurations KOS interface displays by single call
  - novel interfaces, such as navigation via semantic expansion
  - Query expansion for various ranked result query services
  - Term suggestion to assist indexing/annotation
  - More details:
    KOS at your Service: Programmatic Access to Knowledge Organisation Systems http://journals.tdl.org/jodi/article/view/jodi-124/109

SKOS based lightweight semantic services

- Arguably, information retrieval KOS provide a semantic structure at a suitable granularity for the general problem of search and retrieval where fuzzy *aboutness* relationship connects concepts and information resources
- SKOS standard representation, combined with other developments in standard identifiers and service protocols, offers a lightweight approach for a wide variety of annotation, search and browsing oriented applications that don't require first order logic.

Contact Information

Doug Tudhope
School of Computing
University of Glamorgan
Pontypridd CF37 1DL
Wales, UK

dstudhope@glam.ac.uk
http://hypermedia.research.glam.ac.uk/

References

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

CIDOC CRM http://cidoc.ics.forth.gr/

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

STAR Project. http://hypermedia.research.glam.ac.uk/kos/star

Tudhope D., Binding C. 2006. Towards Terminology Services: experiences with a pilot web service thesaurus browser. ASIST Bulletin, 32(5), 6-9, June/July. Available online at http://www.asist.org/Bulletin/Jun-06/tudhope_binding.html

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