Ceri Binding, Hypermedia Research Unit, University of South Wales

5 star data – achieving the 5th star

1 Introduction

Semantic **En**richment **E**nabling **S**ustainability of Ar**cha**eological **L**inks (SENESCHAL) [1] was a 12 month AHRC funded project coordinated by the Hypermedia Research Unit at the University of South Wales. Project collaborators include English Heritage, the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS) and Wales (RCHAMW), Archaeology Data Service (ADS) and others. The project aims included:

- Widening access to key vocabulary resources. National cultural heritage thesauri and vocabularies are used by both national organizations and local authority Historic Environment Records and could potentially act as vocabulary hubs for the Web of Data. Making the terminology more openly accessible as Linked Open Data (LOD) could encourage wider adoption of standard terminology and engender useful community feedback on possible improvements to the existing vocabularies.
- Improving the consistency of existing metadata. This would be achieved by exemplar bulk semantic enrichment operations to align legacy datasets with controlled vocabularies.
- Improving the consistency of future metadata. Better integration of controlled vocabulary resources in the data creation workflow could be facilitated using a suite of interactive web services and associated user interface controls.

As a result of this recently completed project a set of prominent UK archaeological thesauri and vocabularies is now freely available as LOD via http://www.heritagedata.org - together with open source web services and user interface controls.

2 Alignment of legacy data sets to LOD vocabularies

In the course of the SENESCHAL project work it was observed from legacy archaeological datasets the frequent use of free text in data entry fields, leading to simple syntactic anomalies. Minor differences in spelling or punctuation can hinder the alignment of data, preventing wider interoperability. Some errors may not benefit from a simple spell checking approach, as they can form valid words in their own right. Issues empirically identified in archaeological datasets ranged from simple spelling errors to a conscious unilateral attempt to create additional structure, embellishment or description within free text fields. Some examples of these issues are illustrated in the *data value* column of Table 1.

A string similarity algorithm was employed to identify candidate matches by comparing legacy data values to thesaurus concept labels and returning the best scoring matches. The *Levenshtein* edit distance algorithm introduces some controlled flexibility in matching, by measuring the optimal number of character edits required to change one string into another, so accommodating small spelling differences or errors. Any bracketed qualifiers were removed from data values prior to matching. The output of the matching algorithm is converted to a percentage match for display purposes.

Data value	Highest scoring thesaurus term match	Score
AXE FACOTRY	Axe Factory	90%
BOUNDARIES	BOUNDARY	77%
BOUNDARY	BOUNDARY	100%
BUIED SOIL HORIZON	BURIED SOIL HORIZON	97%
CAIRN	CAIRN	100%
CAIRN (POSSIBLE)	CAIRN	100%
CAIRNN	CAIRN	90%
CESS PITT	CESS PIT	94%
CHAMBERED TOM	CHAMBERED TOMB	96%
COMERCIAL	COMMERCIAL	94%
CROFT?	CROFT	90%
CUP-MARKED STONE	CUP MARKED STONE	93%
DICTH	DITCH	80%
ENCLSOURE	ENCLOSURE	88%
EXTRACTION PIT	EXTRACTIVE PIT	85%
EXTRACTIVE PIT	EXTRACTIVE PIT	100%

Table 1 – examples of matching between archaeological data and thesaurus terms using a string similarity algorithm

3 Mapping between vocabularies

One issue that has become more visible since the creation and aggregation of these online resources is that while there is fairly rich intra-thesaurus concepts linkage, there are currently no inter-thesaurus links present. A further related issue is that there are currently very minimal links out to external Linked Data resources. Tim Berners-Lee devised a useful 5 star deployment scheme **Error! Reference source not found.** with which to grade LOD, indicating that the SENESCHAL thesauri currently achieve 4 stars:

★	Data made available on the web - in any format (with an open licence)
★★	As above, but using a machine readable structured data format (e.g. Excel)
★★★	As above, but using non-proprietary structured data formats (e.g. XML)
★★★★	As above, but using W3C open standards (e.g. URIs, RDF & SPARQL)
★★★★★	As above, and also **linking out** to other external LOD

Figure 1 - the 5 star depoyment scheme for LOD

Some of the thesauri converted in the SENESCHAL project actually share a common origin - RCAHMS and RCAHMW each have a separate Monument Type thesaurus, both derived from the original English Heritage Monument Types Thesaurus. English Heritage and RCAHMS also have separate Archaeological Object Types thesauri, derived from a thesaurus originally developed by the Archaeological Objects Working Party. Clearly there is great scope here for some fairly straightforward inter-thesaurus linking of concepts. Making links to other external LOD resources (e.g. Getty Art & Architecture Thesaurus) would constitute achieving the final star in this 5 star scheme.

One of the work packages in the current ARIADNE FP7 project [3] is concerned with the interlinking of archaeological datasets, for the purposes of searching and browsing across an integrated data infrastructure. A logical approach for this work package could be to create suitable links between the

concepts of the various (multilingual) controlled vocabularies associated with the datasets. In matching between thesauri it is necessary to decide on a suitable architecture for mappings. The maximum number of links for many-to-many (M2M) architecture is n^2-n (where n is the number of datasets), for hub architecture it is 2n. For a small project interlinking just 2 or 3 datasets the M2M architecture is satisfactory, for anything above 3 datasets the HUB architecture is probably more appropriate.

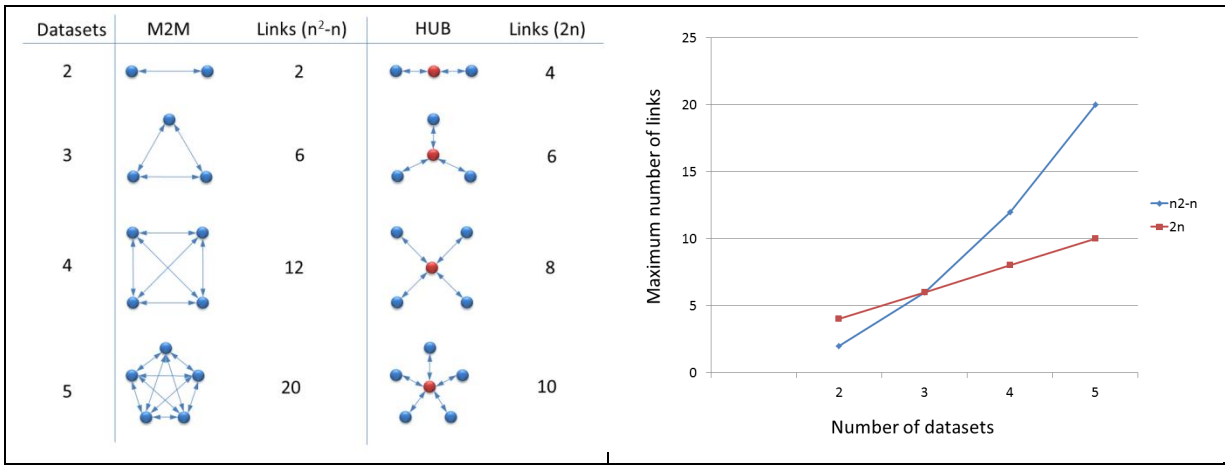

Figure 2 – maximum number of links using M2M and HUB architecture

4 Use of contextual evidence in mapping

Automated tools can assist to an extent, but should be used in conjunction with domain expert mediation to ensure consistent quality of mappings. Results still require manual oversight using other contextual data associated with the concept, as even a 100% match on preferred terms is still only a syntactic match; it does not guarantee a semantic match, as illustrated in Figure 3.

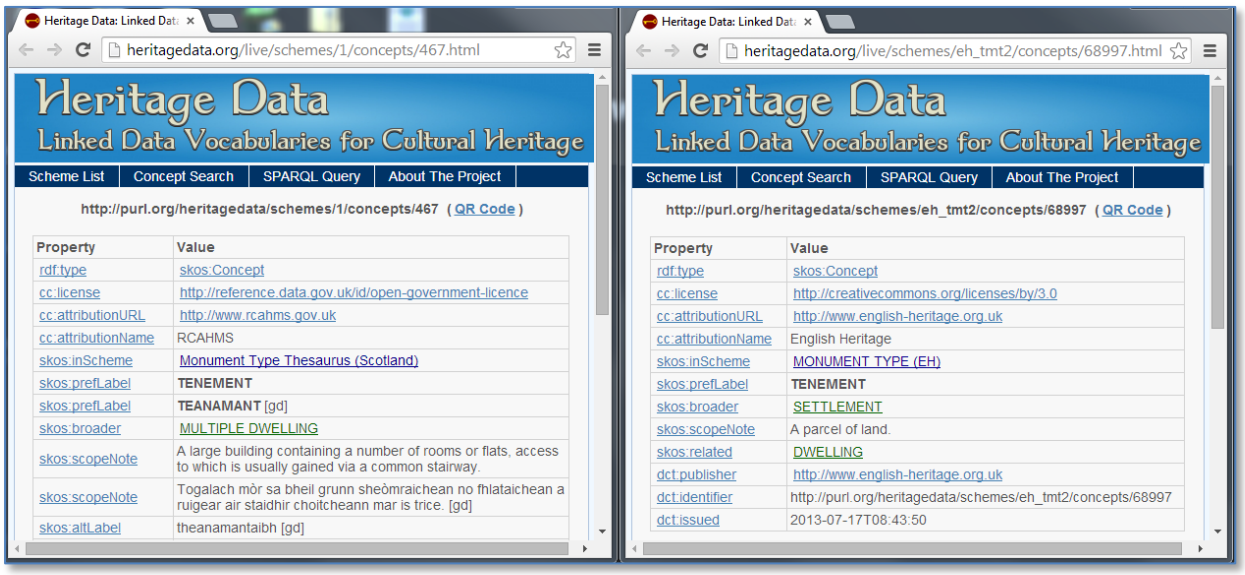

Figure 3 - 100% textual match on preferred label, but further context shows the two concepts are different

The requirement is *concept* alignment, not just *term* alignment, and so additional contextual evidence may be exposed and used by both tools and humans to further qualify a match. For example:

- *Syntactic matching* - may be inexact matching, employing stemming, string matching algorithms (e.g. using the *Levenshtein* edit distance approach as described previously). May need to strip term 'qualifiers', and consider white space, punctuation, capitalisation, case sensitivity etc. Terms may require translation in the case of multilingual terminology.
- *Scope note evidence* – there may be full or partial (or no) overlap in scope between concepts, realistically this contextual evidence requires human oversight. Scope notes may require translation in the case of multilingual terminology.
- *Synonyms* – groups of alternate synonymous terms may help to reinforce the case for a match between two concepts.
- *Hierarchical context* – ancestors and descendants. If a top-down approach is employed there may be existing mappings higher up in the structure that can give additional contextual evidence to a potential match under consideration. The Ontology Alignment Evaluation Initiative (OAEI) [5] 2013 Library Test Case in matching two real world thesauri [6] noted that *"matchers still rely too much on the character string of the labels […] incorrect matches could be prevented […] by taking these higher levels of the hierarchy into account […] We believe that further exploiting this context knowledge could be worthwhile"*.

It is also important to record additional metadata about the mappings being produced, as a new set of mappings constitutes a dataset in its own right and so requires appropriate authorship and licensing information. One approach to this is the use of the VoID vocabulary [7], which may be used to describe linked RDF datasets using the *Linkset* element.

5 Tools and techniques

The issues described above require suitable techniques, methodologies and practical tools to devise mappings between thesaurus concepts. ISO 25964-2:2013 [4] describes approaches for creating mapping relationships between concepts in different vocabularies. It notes the need for caution, stating *"…it is better to have no mapping at all than to establish a misleading one".* Section 14 of the standard discusses some techniques for identifying candidate mappings.

General tools exist for creating mappings between linked data items [8] (e.g. Silk Link Discovery Framework [9]). However the focus of such tools is typically on functionality and automation, they do not necessarily present the user with sufficient contextual data to make an informed academic decision on mappings. In the case of thesaurus to thesaurus mapping it might be useful for instance to compare hierarchical structures side by side, displaying any existing confirmed mapping links between these structures as well as any candidate links. User centred tools or extensions more tailored for the specific task of thesaurus to thesaurus mapping, together with documented methodologies, techniques and approaches could improve the accuracy of the overall process.

6 References

[1] SENESCHAL project [http://hypermedia.research.southwales.ac.uk/kos/SENESCHAL/]
[2] Berners-Lee, Tim. Linked Data – Design Issues [http://www.w3.org/DesignIssues/LinkedData.html]
[3] ARIADNE project [http://www.ariadne-infrastructure.eu/]

[4] ISO 25964-2:2013 Information and documentation — Thesauri and interoperability with other vocabularies -- Part 2: Interoperability with other vocabularies [http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=53658]
[5] Ontology Alignment Evaluation Initiative [http://oaei.ontologymatching.org]
[6] Bernardo Cuenca Grau, Zlatan Dragisic, Kai Eckert, J´erˆome Euzenat,Alfio Ferrara, Roger Granada, Valentina Ivanova, Ernesto Jim´enez-Ruiz, Andreas Oskar Kempf, Patrick Lambrix, Andriy Nikolov, Heiko Paulheim, Dominique Ritze, Franc¸ois Scharffe, Pavel Shvaiko, C´assia Trojahn Ondˇrej Zamazal Results of the Ontology Alignment Evaluation Initiative 2013, pp.29-31 [http://disi.unitn.it/~p2p/OM-2013/oaei13_paper0.pdf]
[7] Keith Alexander, Richard Cyganiak, Michaeil Hausenblad, Jun Zhao. Describing Linked Datasets with the VoID Vocabulary. W3C Interest Group Note (2011) [http://www.w3.org/TR/void/]
[8] References to tools and papers about link generation techniques [http://esw.w3.org/TaskForces/CommunityProjects/LinkingOpenData/EquivalenceMining]
[9] Silk Link Discovery Framework [http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/]