Linking Thesauri - ELSST as a Hub for Social Science Data Terms

by Lorna Balkan and Lucy Bell[1]

Abstract

Without controlled index terms, data retrieval within a data catalogue becomes at best hit and miss. The UK Data Archive manages two thesauri: the multilingual ELSST thesaurus, and the monolingual HASSET thesaurus, from which ELSST is derived. Over the last year, through funding from the UK's Economic and Social Research Council (ESRC), the Archive has developed both of these thesauri, plus their management applications, and created linkages between them. Extending ISO 25964, the Archive has developed a way of mapping its social science thesauri to facilitate cross-national data retrieval. It has also created SKOS formats and is developing a new and innovative application for thesaurus management which combines term visualisation with tree structures. The two thesauri now share a clearly defined, common set of core concepts, but have room for divergence. The new application will allow terms to be promoted to the core set, where they exhibit partial or exact equivalence, or be demoted to 'non-core'. The application allows authorised language equivalents to be added to ELSST terms. Bundled suggestions will also be made for changes to the thesaurus terms or structure, linked to the tree structure. This paper describes the new application and the processes used within the Archive for thesaurus management.

Keywords: : Thesaurus management, thesaurus mapping, SKOS, multilingual thesauri

Introduction

Traditionally, thesauri have held a key role in data archiving as aids for searching and browsing. More recently, they are finding new applications in the world of Linked Data and the Semantic Web. The UK Data Archive[2] has wide expertise in thesaurus development. It has been developing the monolingual social science thesaurus Humanities and Social Science Electronic Thesaurus (HASSET[3]) for over forty years, and for the last 10 years has been managing it in tandem with the related multilingual European Language Social Science Thesaurus (ELSST[4]). Recently it has received funding from the UK's Economic and Social Research Council (ESRC) for a 5-year project (2012-2017), CESSDA-ELSST[5], to update and develop both thesauri, and implement

> Thesauri and other controlled vocabularies also play a vital role in Linked Data.

a new management application and processes. This paper describes work to date, challenges faced, and solutions adopted.

Background

The work reported in this paper was motivated by two main concerns. The first was to update the thesauri managed by the UK Data Archive so that they can be exploited and shared more easily. The second was to find a way of managing them more efficiently.

The changing role of thesauri

Thesauri are a type of controlled vocabulary. Controlled vocabularies "mandate the use of predefined, authorized terms that have been preselected by the designer of a vocabulary, in contrast to natural language vocabularies, where there is no restriction on vocabulary" (Andritsos and Keilty, 2014). Specifically, thesauri belong to the class of Knowledge Organization Systems (KOSs) that are "controlled vocabularies, which are organized and structured via different types of semantic relationships" (Golub and Tudhope, 2009).

Traditionally thesauri have been used to support human-mediated access to information. Increasingly, they and other controlled vocabularies are being used for machine-to-machine communication and to underpin web services (Dextre Clarke and Zeng, 2012). Web services include terminology services, defined as "a group of abstract services, presenting and applying vocabularies, their member concepts, terms and relationships, describing the meaning of terms and facilitating semantic interoperability. This is done for purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc." (Tudhope et al., 2006). Examples of terminology services include the NERC vocabulary server[6], UMLS[7] and the BioPortal[8]. Terminology services may be used in isolation or in combination with a wide range of other web services (see Golub and Tudhope (2009) for a discussion of use cases).

Thesauri and other controlled vocabularies also play a vital role in Linked Data. Linked Data is "an approach to data integration that employs ontologies, terminologies, Uniform Resource Identifiers (URIs) and the Resource Description Framework (RDF[9]) to connect pieces of data, information and knowledge on the Semantic Web" (Marshall et al., 2012). Vocabularies published as Linked Data can be linked to other vocabularies, allowing databases indexed with one vocabulary to be searched using another (Méndez and Greenberg, 2012). Some researchers, including Shiri, also predict that controlled vocabularies will have a role to play in the new Big Data landscape: "General purpose and domain-specific controlled vocabularies published as linked open vocabularies can not only be used to organize and represent structured data such as linked data repositories and semantic web applications, they can also be used to index, organize and analyze unstructured textual information that exists in several big data sources."(Shiri, 2014).

To enable thesauri and other vocabularies to be fully exploited requires them to be interoperable. Recent years have seen the emergence of new standards designed to promote interoperability. Amongst the most important is the publication of ISO 25964[10], the new international guidelines on thesaurus construction. It is in two parts. Part 1 covers the construction of monolingual and multilingual thesauri, while Part 2 is concerned with interoperability and mapping between different vocabulary types, including thesauri. Crucially ISO 24964 contains an explicit data model that clearly distinguishes between concepts and the terms used to represent the concepts. Dextre Clarke and Zeng (2012) argues that "To perform on the Semantic Web, computer software needs an explicit data model that distinguishes between terms and concepts".

Interoperability also requires common encoding schemes. Simple Knowledge Organization System (SKOS[11]), which uses RDF, is emerging as the preferred standard in which to encode thesauri and other KOSs. Many prominent thesauri have already been converted to SKOS and made available as Linked Data (see for example AGROVOC (Caracciolo et al., 2013), EUROVOC[12] and GEMET[13]).

Key goals of the CESSDA-ELSST project were, therefore, to convert the thesauri from a term-based to a concept-based model, following ISO 25964, and to make them available as SKOS-based Linked Data.

Thesauri at the UK Data Archive

Thesaurus development has been a key activity of the UK Data Archive since its inception. HASSET is the in-house thesaurus of the UK Data Service, and is used to index and search the Service's data collection, which, with over 6,000 datasets, is the largest social science data collection in the UK. HASSET was originally derived from the UNESCO thesaurus, but has been developed in-house at the Archive for over 40 years. It currently contains 4,743 preferred terms, and is available to external users under license.

ELSST is a multilingual thesaurus that is used in the CESSDA data portal. The Consortium of European Social Science Data Archives (CESSDA) is a body that promotes the acquisition, archiving and distribution of electronic data for social science teaching and research in Europe.

ELSST began life in 2000 as part of the EU-funded Language Independent Metadata Browsing of European Resources (LIMBER[14]) project, with the aim of enhancing cross-border data discovery and utilization. ELSST has been developed further over the years through additional funding from the ESRC and the University of Essex and via other EU-funded projects including the Multilingual Access to Data Infrastructures of the European Research Area (MADIERA[15]) and CESSDA-Preparatory Phase Project (CESSDA-PPP[16]). It is currently available in 9 languages (Danish, English, Finnish, French, German, Greek, Norwegian, Spanish, Swedish) with more in progress, including Czech, Lithuanian and Romanian. It was originally derived from HASSET and English continues to be the source language. It currently contains 3,286 English preferred terms. Like HASSET, it is available to external users under license.

Although the two thesauri have much in common (almost all ELSST terms are also in HASSET), historically, they have been managed on different platforms. Not only has this led to much duplication of effort, since often the same information has to be entered in two different places, it has led to the two thesauri diverging without this being desired or obvious. One of the main aims of the CESSDA-ELSST project was, therefore, to bring the two thesauri together onto one platform, and possibly merge them. It was expected that this would result in resource efficiencies and improved quality assurance.

The design of the new system had to take account of the fact that the two thesauri work to differing time scales. HASSET is constantly updated so that indexers within the UK Data Archive and members of the general public may browse the current version at all times, while ELSST is moving towards an annual version release. Another difference concerns validation procedures – HASSET terms are validated in-house, while additions or changes to ELSST source language terms require the approval of its international translators committee, of which the UK Data Archive is the chair. (Translations of source terms are, however, the sole responsibility of translators or their institutions.)

The conversion of both thesauri to a concept-based model following ISO 25964 was expected not only to improve their interoperability but to bring efficiency gains for their management. As Dextre Clarke and Zeng (2012) points out: "Benefits of adopting the [data] model include easier implementation by computers, consistency enforced in thesaurus construction and mapping, greater interoperability between thesauri and with other vocabularies, and enhanced performance at all stages of the thesaurus through development, management, and exchange." ISO 25964-2 also proved helpful to articulating the relationship between HASSET and ELSST.

Thesaurus development work
Development work on the thesauri was broken into a number of sub-tasks, some of which overlapped, and some of which are still ongoing. The main ones are described below.

Revising and updating terms and structures
A major and ongoing part of the project is to review and update the terms and structure of both thesauri. As part of this work, a thorough review of the top terms is being undertaken, with a view to reducing them in number. Reducing the number of top terms (currently 298 in HASSET, 218 in ELSST) will enhance the usefulness of the thesauri as a browsing aid.

Another focus of this sub-task is to review the terms themselves, removing redundancy where it has occurred, and ensuring that terms are up-to-date. Up-to-date terms are of particular importance for automatic indexing. HASSET has already been used for automatic indexing (El Haj et al., 2013), and further experiments are planned in future.

As terms are revised or added, scope notes are added where possible. These serve to define the semantic boundaries of the term and are useful not just to the thesaurus developers (in particular to ELSST developers when looking for a translation of an English term), but also to users of the thesaurus. ISO 25964 will be followed wherever possible.

From term-based to concept-based model
As mentioned above, an important part of the CESSDA-ELSST project was to move the thesauri from a term-based to a concept-based model, following ISO 25964. Thus preferred terms become labels for concepts, and in a multilingual thesaurus like ELSST, different language versions of preferred terms are just alternative labels for the same concept. Each concept has its own GUID (Globally Unique IDentifier). In a term-based thesaurus semantic relationships are established between the terms themselves. In a concept-based thesaurus certain semantic relationships (e.g. hierarchical and associative relationships) are established between the concepts, and others (e.g. equivalence) between terms (Dextre Clarke and Zeng, 2012). The concept-based model has advantages over the term-based model, not just for thesaurus management, but also for indexing. Documents are associated with concepts, not terms, thus changes involving preferred and non-preferred terms do not impact on indexing (Pastor-Sanchez et al., 2009).

SKOS
In order to promote interoperability, converting both thesauri to Simple Knowledge Organization System (SKOS) was a crucial goal of the project.

The main objective of SKOS is to enable the easy publication of KOSs for the Semantic Web. Like ISO 25964, SKOS is concept-based, and, according to the ISO 25964 homepage, care has been taken by ISO 25964 developers to maintain compatability with SKOS (see ISO 25964).

ELSST is being converted to SKOS as part of the CESSDA-ELSST project – HASSET was already converted to SKOS during the SKOS-HASSET[17] project (Bell, 2013). The SKOS version of HASSET is available to external users under license, and the SKOS version of ELSST will be available in early 2015. The SKOS versions of both thesauri are implemented using GUIDs and BrightStarDB[18] for the triple stores, and published via Pubby[19], which provides a browseable, meaningful view of the thesaurus (Bell, 2012).

Defining the basics
An important prerequisite to designing the new thesaurus management system was to establish all the possible elements and relationships in each thesaurus. The three basic relationships within a thesaurus are as defined in ISO 25964: equivalence, which holds between a preferred term (PT) and a non-preferred term or Use For (UF); hierarchical, which holds between a broader term (BT) and a narrower term (NT); and associative, which holds between related terms (RTs). In the legacy systems, scope notes in both thesauri included information about meaning as well as usage and history of a term. In the new system, this information will be separated into different types of notes, namely scope notes, use notes and history notes respectively. Historically, scope notes in ELSST have also included 'translation notes' that describe any difference in meaning between the English source term and its equivalent in another language. In the new model, translation notes will be also recorded in a separate field. These changes were introduced to make it easier for both users and developers to distinguish the different types of information, and additionally, to help developers keep track of the differences between HASSET and ELSST concepts.

Axioms and constraints
Once the basic elements of the thesauri were established, the next step was to define the axioms and constraints that hold among concepts within and between each thesaurus. An example of the former type of constraint is the requirement that a term may be a Use For (UF) to only one PT within the thesaurus (or same language version of the thesaurus, in the case of ELSST). Inter-thesaural axioms and constraints were refined and updated as a result of the thesaurus alignment exercise, described below.

Thesaurus alignment exercise
The aim of the alignment exercise was to see whether the two thesauri could be merged. All the terms and relationships that were in ELSST, not HASSET, were examined, and the differences between the two thesauri resolved wherever possible. Further alignment work will look at the terms and relationships that are in HASSET, not ELSST, to see whether they should be brought into ELSST.

Results from the first part of the alignment exercise suggested that, instead of forcing the two thesauri to merge, their common set of core concepts should be kept identical wherever possible, but allowed to diverge in clearly defined ways. Rather than being seen in terms of merging, their relationship can best be described in terms of a mapping. In this way, both thesauri can retain their integrity and identity.

Mapping ELSST to HASSET

ELSST and HASSET will have a set of shared or 'core' concepts. Non-core concepts will also be possible in both thesauri. A mapping relationship, based on the equivalence relationship defined in ISO 25964-2, will hold between all core concepts.

ISO 25964-2 defines three types of mapping between thesauri: equivalence, hierarchical, and associative. An equivalence mapping is established when matching concepts are found in two or more different vocabularies potentially with different preferred term labels. Equivalence may be 'simple' (when the two thesauri contain concepts that are identical in scope), or 'compound' (where a concept represented in one vocabulary with just one preferred term may be represented in another vocabulary by a combination of two or more concepts/terms). Simple equivalence mappings may also be either exact or inexact. Exact equivalence arises when "the concepts can be used interchangeably across all the applications that can be envisaged for the mapping" (ISO 25964-2, Section 11.2). Inexact equivalence, by contrast, arises when the concepts are equivalent in some contexts but not others, or where concepts have overlapping scopes or small differences of connotation.

The equivalence mapping between ELSST and HASSET is defined as follows. Core concepts will exhibit either 'exact' or 'close' equivalence. For 'exact equivalence' to hold, the concepts must have the same preferred terms (PTs), broader terms (BTs), scope note and scope note source. 'Close equivalents' will only be required to have the same PTs and BTs. In both cases, all other associated metadata may differ, including: Use For (UFs), narrower terms (NTs), related terms (RTs), use notes, etc.

Note that these definitions are stricter than the definition of ISO 25964-2 equivalence mapping, since they demand identity at structural and term/linguistic level, as well as semantic level. The ISO 25964-2 definition of equivalence, by contrast, is entirely semantic. The ELSST/HASSET 'exact equivalence' corresponds to 'exact simple equivalence' in ISO 25964. We may expect any difference between the scope notes of 'close equivalents' in ELSST and HASSET also to be small enough for 'exact simple equivalence' to hold in ISO terms. That is to say, the difference in meaning between 'exact' and 'close' ELSST/HASSET equivalents will have no significant impact on information retrieval.

Not all restrictions on the relationship between core concepts in the two thesauri can be captured by axioms and constraints (for example, when scope notes and other metadata may differ), and developers will rely on the reporting functions of the thesaurus management system to keep track of all the differences between the two thesauri. Work continues on when shared concepts may differ.

Thesaurus management system

All ELSST partners contributed to the requirements gathering of the new thesaurus management system. They were also invited to give feedback on the first prototype, and their comments were used to inform the final version. The system is now complete and has received excellent feedback from the testing by translators.
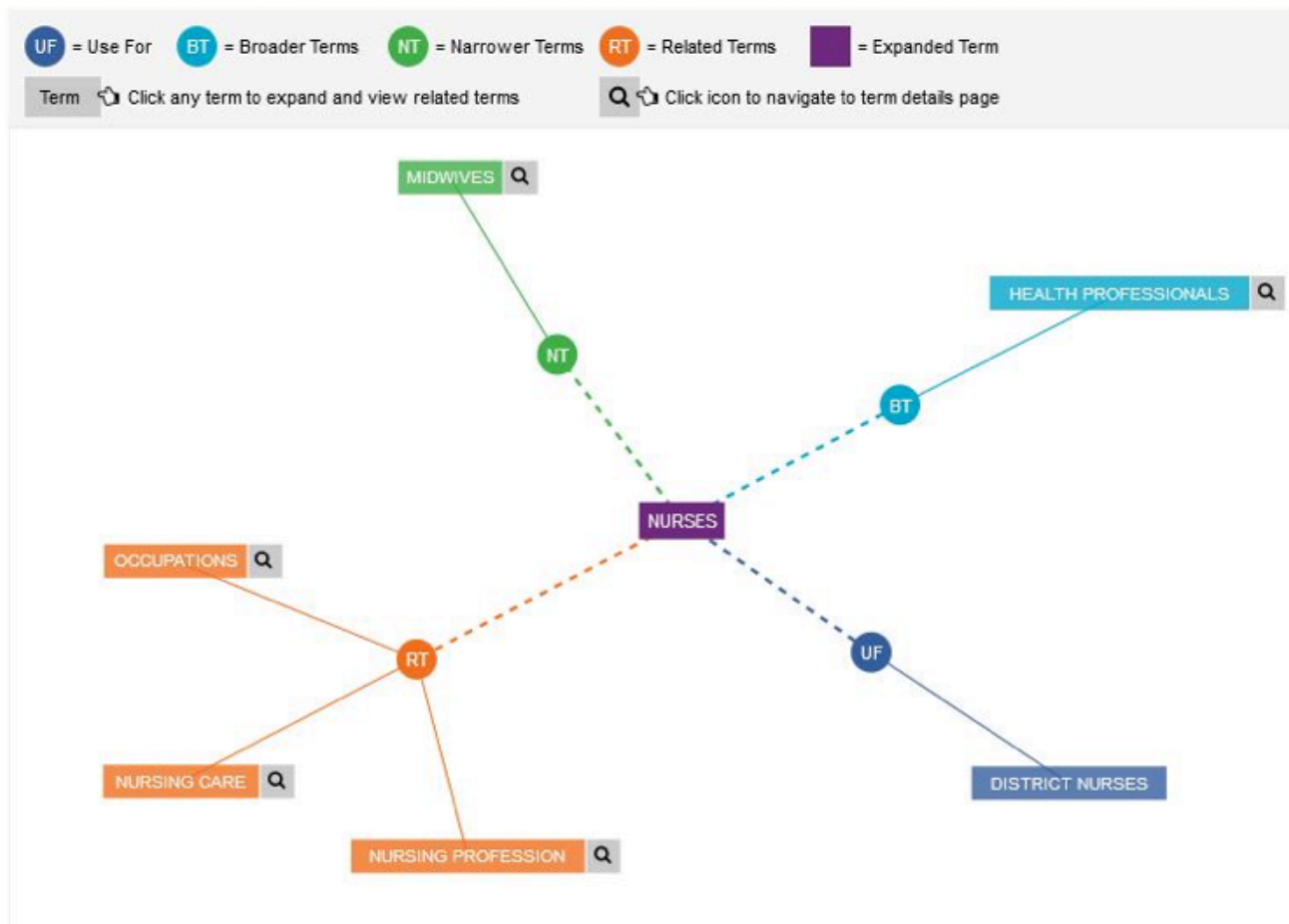

Figure 1 Visual graph view of the concept NURSES

The central feature of the new thesaurus management system is that the two thesauri now share the same database. This is essential to managing them in an efficient manner. However, the two thesauri will have separate user interfaces. HASSET terms will be linked to the studies at the UK Data Service that have been indexed with them, while ELSST will provide a link to its multilingual equivalents. The user interfaces otherwise have identical features.

A novel feature of both the user and management interfaces is the implementation of a visualisation tool for navigation, in addition to the traditional tree structures. As data are becoming ever bigger and more complex, data visualisation is becoming increasingly popular as an alternative and more user-friendly way of viewing data. Data visualisation of knowledge organization systems in particular is a lively research topic (see for example Katifori et al., 2007 for an overview) and has been the focus of many recent workshops (see for example Slavic et al. 2013). The visualisation solution adopted for the new thesaurus management system is an interactive tool which presents concepts to the user in a colour-coded, expandable graph (see Figure 1).

A well-designed management interface is key to the smooth management of both thesauri. Management permissions are controlled via Shibboleth[20], and authorised users can, as appropriate, suggest, discuss, and implement changes or translations, in the relevant thesaurus. Core concepts can be 'demoted' to non-core concepts in either ELSST or HASSET, and conversely, non-core concepts in HASSET can be 'promoted' to core concepts.

An innovative feature is the suggestions area where a number of proposed changes can be bundled together as one suggestion, since a change to one term is often associated with a change to another. The bundled suggestions are also linked to the tree structure. The suggestions area also provides a place where changes to terms can be discussed and agreed with external partners.

Since it is essential for the UK Data Service developers to keep track of the differences between HASSET and ELSST core concepts, and since not all of these can be captured by axioms and constraints, reporting functions play a vital role. To facilitate this, terms will be deprecated, but never deleted.

Technologies used for implementing the system include AJAX scripting for the user interface, to minimise post backs and maximise the user experience, LINQ to XML, and SOLR.

Conclusion and future directions

The CESSDA-ELSST project has made good progress to date. Development work in this first phase has concentrated on converting the thesauri to a concept-based model, creating a SKOS version of ELSST, aligning the two thesauri, and defining the relationship between them. These in turn have enabled the design and implementation of the new thesaurus management system. Archive staff and ELSST translators are now keen to use the system in their everyday work. Future work will include refining the structure, and updating the content of the two thesauri further, as well as possible enhancements to the management system, based on user feedback.

The project is expected to produce significant gains for both end users and developers. UK developers will benefit from a new and improved management application, which will allow them to keep track of the differences between the two thesauri and thus save time and effort. They will also have a more user-friendly platform for managing suggestions and changes to ELSST with their international partners.

End users will benefit from the updated content and structure of both thesauri, and improved access to them via the new user interfaces. Compliance with ISO 25964 and conversion to SKOS will not only enhance ELSST's status as a hub for social science data terms but open it up to the world of the Linked Data and the Semantic Web.

Acknowledgements

References

Andritsos, Periklis and Keilty, Patrick (2014) Level-Wise Exploration of Linked and Big Data Guided by Controlled Vocabularies and Folksonomies. Advances In Classification Research Online. 24(1). doi:10.7152/acro.v24i1.14670 (Available at http://journals.lib.washington.edu/index.php/acro/article/view/14670/12310)

Bell, Darren (2012) From Tuples to Triples: applying SKOS to HASSET – a technical overview. SKOS-HASSET blog. (Available at http://hassetukda.wordpress.com/2012/12/20/from-tuples-to-triples-applying-skos-to-hasset-a-technical-overview/)

Bell, Lucy (2013) SKOS-HASSET: Help for users. (Available at http://www.data-archive.ac.uk/media/393116/skos-hasset-help.pdf)

Caracciolo, Caterina; Stellato, Armando; Morshed, Ahsan; Johannsen, Gudrun; Rajbhandari, Sachit; Jaques, Yves; and Keizer, Johannes (2013) The AGROVOC Linked Dataset. Semantic Web, 2013 4(3) pp. 341-348. (Available at http://eprints.rclis.org/20648/)

Dextre Clarke, Stella and Zeng, Marcia Lei (2012) From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modeling. Information Standards Quarterly. 24(1). (Available at http://www.niso.org/publications/isq/2012/v24no1/clarke/)

El-Haj, Mahmoud; Balkan, Lorna; Barbalet, Suzanne; Bell, Lucy and Shepherdson, John (2013) An Experiment in Automatic Indexing Using the HASSET Thesaurus. The 5th Computer Science and Electronic Engineering Conference (CEEC'13), IEEE Xplore. 17-18 September 2013, University of Essex. doi: 10.1109/CEEC.2013.6659437 (Available at http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6659437)

Golub, Koraljka and Tudhope, Douglas (2009) Terminology Registry Scoping Study (TRSS) Final Report. (Available at http://www.jisc.ac.uk/media/documents/programmes/sharedservices/trss-report-final.pdf)

Katifori, Akrivi; Halatsis, Constantin; Lepouras, George; Vassilakis, Costas and Giannopoulou, Eugenia (2007) Ontology visualization methods – a survey. ACM Computing Surveys, 39(4). (Available at http://dl.acm.org/citation.cfm?id=1287621)

Marshall, M. Scott; Boyce, Richard; Deus, Helena F.; Zhao, Jun; Willighagen, Egon L.; Samwald, Matthias; Pichler, Elgar; Hajagos, Janos; Prud'hommeaux, Eric and Stephens, Susie

(2012) Emerging practices for mapping and linking life sciences data using RDF - a case series. Web Semantics: Science, Services and Agents on the World Wide Web. Volume 14. July 2012. pp. 2–13. (Available at http://www.researchgate.net/publication/253234505_Emerging_practices_for_mapping_and_linking_life_sciences_data_using_RDF__A_case_series)

Méndez, Eva and Greenberg, Jane (2012) Linked data for open vocabularies and Hive's global framework, El Profesional de la Informacion. May/June 2012. 21(3). (Available at http://www.elprofesionaldelainformacion.com/contenidos/2012/mayo/03_eng.pdf)

Pastor-Sanchez, Juan-Antonio; Martínez Mendez, Francisco Javier and Rodríguez-Muñoz, José Vicente (2009) Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. Information Research. 14(4). December 2009. (Available at http://files.eric.ed.gov/fulltext/EJ869364.pdf)

Shiri, Ali (2014) Linked Data Meets Big Data: A Knowledge Organization Systems Perspective, Advances In Classification Research Online, 24(1). doi:10.7152/acro.v24i1.14672. (Available at http://journals.lib.washington.edu/index.php/acro/article/view/14672/12312)

Slavic, Aida; Akdag Saha, Almila and Davies, Sylvie (eds.) (2013): Classification and visualization: interfaces to knowledge: proceedings of the international UDC Seminar, 24-25 October 2013, The Hague, The Netherlands. Ergon Verlag, Würzburg.

Tudhope, Douglas; Koch, Traugott and Heery, Rachel (2006) Terminology services and technology: JISC state of the art review. (Available at http://www.ukoln.ac.uk/terminology/JISC-review2006.html)

Notes

1. Both authors work at the UK Data Archive, University of Essex. Lorna Balkan is CESSDA-ELSST Co-ordination Officer. She can be contacted at balka@essex.ac.uk. Lucy Bell is Functional Director, Data Access. She can be contacted at lajbell@essex.ac.uk. This paper was presented in the 'Harmonization, Thesauri and Indexing' session at IASSIST 2014
2. http://www.data-archive.ac.uk
3. http://hasset.ukdataservice.ac.uk
4. http://elsst.ukdataservice.ac.uk
5. http://ukdataservice.ac.uk/about-us/projects/cessda-elsst/details.aspx
6. http://www.bodc.ac.uk/products/web_services/vocab/
7. http://www.nlm.nih.gov/research/umls/
8. http://bioportal.bioontology.org/
9. http://www.w3.org/RDF/
10. http://www.niso.org/schemas/iso25964/
11. http://www.w3.org/2004/02/skos/
12. http://open-data.europa.eu/en/data/dataset/eurovoc
13. http://www.eionet.europa.eu/gemet/exports/en/rdf/
14. http://www.data-archive.ac.uk/about/projects/limber
15. http://www.data-archive.ac.uk/about/projects/madiera
16. http://www.data-archive.ac.uk/about/projects/cessda-ppp
17. http://www.data-archive.ac.uk/find/our-projects/skos-hasset
18. http://brightstardb.com/
19. http://www.w3.org/2001/sw/wiki/Pubby
20. https://shibboleth.net/