Linking thesauri: ELSST as a hub for social science data terms

Lorna Balkan

CESSDA Thesaurus Coordination Officer
UK Data Archive

IASSIST 2014, Toronto

4 June 2014

UK Data Service

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

Overview

- The UK Data Archive manages two related thesauri:
    - ELSST -   Multilingual European Language Social Science Thesaurus
    - HASSET - Humanities and Social Science Electronic Thesaurus
- received funding from UK's ESRC for CESSDA ELSST project (October 2012 to September 2017) to:
    - **develop both thesauri**
    - redevelop **thesaurus management applications**
    - streamline **thesaurus management processes**
- progress so far

Background

- HASSET and ELSST have, historically, been managed separately, using different administrative interfaces

- This has created problems:

  - it is resource-intensive and inefficient to have to enter the same information in two places

  - maintaining two separate structures creates the capacity for the two thesauri to diverge – or for errors to be entered – without this being obvious

UK Data Service

Thesaurus development: aims

- align two thesauri as far as possible; initial hypothesis – two thesauri can be merged

- update and revise thesauri

- make compliant (as far as possible) with new international thesaurus guidelines for monolingual and multilingual thesauri - ISO 25964, to promote interoperability

- create SKOS version of both thesauri, to make available for linked data

ISO 25964

- ISO 25964 = the international standard for thesauri and interoperability with other vocabularies
- Part 1 Thesauri for information retrieval:
  - covers all aspects of developing a thesaurus, monolingual or multilingual
- Part 2 Interoperability with other vocabularies
  - explains how to map between the concepts in different KOS
- concept-based, rather than term-based thesaurus model

Thesaurus development – work so far

- moved from term-based to concept-based thesaurus model – (following ISO 25964); preferred terms now labels for concepts

- defined which thesaurus elements/relations are required:
    - PTs, UFs, BTs, NTs, UFs
    - Notes: Previous SNs now split into SNs, Use notes, history note, translation notes (ELSST), edit notes

- defined axioms and constraints

- initial alignment work: identified and corrected inconsistencies between the two thesauri at term level and relationship level

Thesaurus development: ongoing work

- review of hierarchical structure including
    - use of polyhierarchies
    - review of top terms
- updating terms
    - removing redundancy of terms
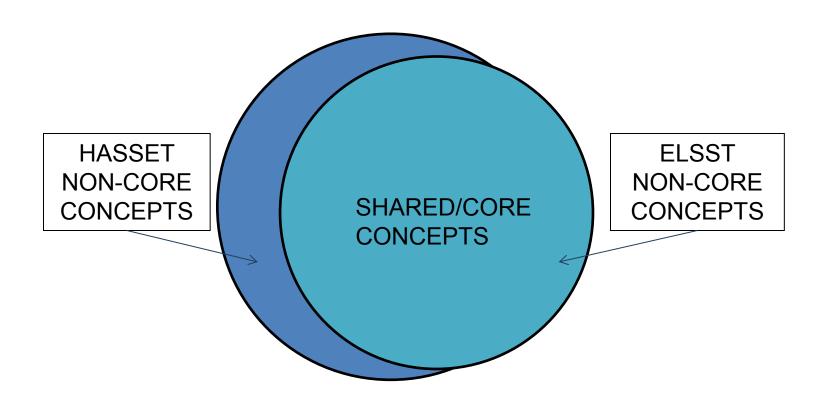   - adding SNs
- further thesaurus alignment work

HASSET/ELSST relationship: overview

- each may consist of shared ('core') concepts and non-core concepts

- ELSST contains internationally applicable concepts only

- HASSET also contains culture-specific concepts (e.g. BARRISTERS) and other non-core concepts and hierarchies (e.g. the GEOGRAPHICAL AREAS AND COUNTRIES hierarchy)

- differing time scales:

  - HASSET is constantly updated so that indexers within the UKDA and members of the general public may browse the current version at all times

  - ELSST moving to annual release, following international approval of new or changed terms

HASSET/ELSST relationship

HASSET
NON-CORE
CONCEPTS

SHARED/CORE
CONCEPTS

ELSST
NON-CORE
CONCEPTS

Thesaurus alignment exercise

- aim to see if feasible to merge ELSST and HASSET, i.e. to consider ELSST a true subset of HASSET

- looked at all concepts, terms and relations that were in ELSST, not HASSET (e.g. concepts that had different BTs in ELSST and HASSET) and resolved where possible

- conclusion - not possible to merge the two thesauri completely

- instead, should aim for alignment as much as possible but, where not, should aim to make the relationships between ELSST and HASSET concepts clear

HASSET/ELSST relationship defined

- rather than being merged, HASSET and ELSST will be mapped – see ISO 25964-2

- allows each thesaurus to retain own identity and integrity – with some rules

- enables the two thesauri to share a clearly defined, common set of shared/"core" concepts while allowing room for divergence

Mapping in ISO 25964-2: overview

- mappings are semantic – they are established between concepts, as represented by their preferred terms, and not between structures

- structure may, however, be taken into account in determining whether a mapping between concepts can be established

Mapping in ISO 25964-2: types

- basic mapping types:
  - Equivalence
    - Simple
    - Compound
  - Hierarchical
    - Broader
    - Narrower
  - Associative

- exact vs. inexact: applies to simple equivalence only – "inexact" equivalence arises when the concepts are equivalent in some contexts but not others, or where concepts have overlapping scopes or small differences of connotation

Mapping relationship between shared/core concepts in HASSET and ELSST

- Two types of equivalence – exact and close
- **"exact equivalent"** must have the same:
  - Preferred Term
  - BTs
  - Scope note & scope note source
- **"close equivalent"** must only have the same:
  - Preferred Term
  - BTs
- other associated metadata may differ, including:
  - non-preferred terms (UFs)
  - NTs
  - RTs

UK Data Service

ELSST → HASSET equivalence versus ISO 25964-2 equivalence

- ELSST → HASSET equivalence has structural constraints (must have same BTs), so can be seen as narrower (and more restrictive) case of the **simple equivalence** as defined by ISO 25964-2. Also requires identity of preferred labels

- **"exact equivalence"** in HASSET/ELSST same as **exact simple equivalence** in ISO 25964-2

- **"close equivalence"** in HASSET/ELSST same as **inexact simple equivalence** in ISO 25964-2

Some implications for development of the two thesauri

- core concepts must be appropriate to all languages
- core BT structure must be appropriate to all languages
- workflow: core BTs must be added at same time – need time to negotiate with partners
- it is important to keep track of differences between core concepts in the two thesauri for thesaurus management purposes – reporting functions important

Thesaurus alignment - outstanding questions

- Coverage: what about concepts that are not currently in ELSST that are not UK-specific – should they be added? Consultation required with ELSST partners

- How and when shared concepts may differ in:
    - o SN
    - o UFs
    - o NTs
    - o RTs

- ELSST -> HASSET mapping model for other local extensions?

- How will mapping be represented/implemented?

UK Data Service

Applying SKOS to HASSET and ELSST-1

What is SKOS?

- Simple Knowledge Organization System

- encodes KOS in a standardised way to make their structures comparable – promotes interoperability

- developed in close association with **ISO 25964**

- set of RDF predicates to describe relationships between thesaurus terms

Applying SKOS to HASSET and ELSST-2

- SKOS already applied to HASSET during Jisc-funded SKOS HASSET project – see

  lod.data-archive.ac.uk/skoshasset/

- soon to be applied to ELSST

- persistence via GUIDs (Globally Unique IDentifiers)

- version control

- use Pubby to publish our SKOS

- use BrightStarDB for the Triple Store

UK Data Service

Example of SKOS-HASSET

```xml
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#" xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
 xmlns:dc="http://purl.org/dc/terms/" xmlns:skos="http://www.w3.org/2004/02/skos/core#"
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
 <skos:ConceptScheme rdf:about="http://tempuri/HASSET/">
   <dc:title>Test SKOS-HASSET Thesaurus</dc:title>
   <rdfs:label>Test SKOS-HASSET Thesaurus</rdfs:label>
   <skos:hasTopConcept rdf:resource="http://tempuri/HASSET/006D08D9-1238-4436-B74B-AB7E0C684CFD" />
   <skos:hasTopConcept rdf:resource="http://tempuri/HASSET/FE276908-AB56-4786-8FCA-5B09E4C9B16C" />
   <skos:hasTopConcept rdf:resource="http://tempuri/HASSET/FFDD3FC3-BC29-42C3-9CE5-CE625F80B9DC" />
 </skos:ConceptScheme>
 <skos:Concept rdf:about="http://tempuri/HASSET/003D1273-6446-4B99-BE8A-C7E9AB56AF71">
   <skos:altLabel xml:lang="en">CRASH HELMETS</skos:altLabel>
   <skos:altLabel xml:lang="en">CYCLE HELMETS</skos:altLabel>
   <skos:broader rdf:resource="http://tempuri/HASSET/EC52047A-6C1F-46F2-9798-2EB24BC7F0A0" />
   <skos:inScheme rdf:resource="http://tempuri/HASSET/" />
   <skos:prefLabel xml:lang="en">PROTECTIVE HEADWEAR</skos:prefLabel>
   <skos:related rdf:resource="http://tempuri/HASSET/7419A454-61B8-42A0-B4D8-4590AFE92FF4" />
   <skos:related rdf:resource="http://tempuri/HASSET/EAE3E8B1-F5BB-432B-BFAE-1E7457FC2DC9" />
 </skos:Concept>
 <skos:Concept rdf:about="http://tempuri/HASSET/0044108B-FD96-4691-AAFC-E47426E3E5BD">
   <skos:altLabel xml:lang="en">FLATS</skos:altLabel>
   <skos:broader rdf:resource="http://tempuri/HASSET/62BA7106-D4B7-4B81-A21E-2D208FEDCDF3" />
   <skos:broader rdf:resource="http://tempuri/HASSET/67E5FC17-9CA7-469A-B3D1-D9372C59DA0E" />
   <skos:inScheme rdf:resource="http://tempuri/HASSET/" />
   <skos:narrower rdf:resource="http://tempuri/HASSET/08E3FB58-6D22-47D4-8469-FB06A42675CD" />
   <skos:prefLabel xml:lang="en">APARTMENTS</skos:prefLabel>
   <skos:related rdf:resource="http://tempuri/HASSET/927EDB93-87E1-4978-AD39-F3D9E1F00403" />
   <skos:related rdf:resource="http://tempuri/HASSET/CA4BBEB9-7C84-4574-8BC0-DC8178DE595E" />
 </skos:Concept>
</rdf:RDF>
```

Applying SKOS to HASSET and ELSST: issues

- consistency checks / "loose terms"
- question of dereferencing concepts
  - SKOS-HASSET is machine-actionable (see www.linksailor.com)
  - how to deliver meaningful web content for, say, *skosconcept:162*?
  - solution: browseable, meaningful thesaurus view www.data-archive.ac.uk/find/hasset-thesaurus/hasset-browser

UK Data Service

Thesaurus management system (TMS)

- single management system for both thesauri
- updated and improved **user-facing pages**
- improved **management interface**

UK Data Service

TMS user-facing interface

- separate interfaces for two thesauri

- supports search and browse functions

- term visualisation and tree structure

- multilingual (ELSST)

- allows linking to datasets (HASSET)

- much of the user interface has been implemented using AJAX scripting to minimise post backs and maximise the user experience, and using restricting lists in real time

UK Data Service

TMS user-facing interface: ELSST homepage

TMS user-facing interface: Thesaurus search

TMS user-facing interface: View term (tree)

TMS user-facing interface: View term (visual graph)

TMS management interface

- enables bundled suggestions to be made for changes to the thesaurus terms or structure, also linked to the tree structure - LINQ to XML

- allows terms to be promoted to the core set, or be demoted to 'non-core'

- enables authorised language equivalents to be added to ELSST terms

- keeps track of changes – terms never deleted

- keeps track of differences between two thesauri

- allows linking to indexing software (HASSET)

TMS management interface: Add suggestion

Thesaurus management process

- single workflow for concept management
- need to share suggestions with colleagues inside and outside the Archive
- fulfils two different time-based requirements

Summary and conclusion

- thesaurus development well under way
- increased interoperability through concept-based model, ISO compliance, and conversion to SKOS
- improved cross-border searching with clearly defined mappings between the two thesauri
- thesauri alignment should reduce time and effort to maintain two separate thesauri
- more user-friendly user-facing interface
- innovative thesaurus management system – should enhance communication

Obtaining HASSET and ELSST

- both free to browse:
  - HASSET:

    www.data-archive.ac.uk/find/hasset-thesaurus
  - ELSST: http://elsst.esds.ac.uk

- licence available to use or adapt thesauri for own organisation:
  - help@ukdataservice.ac.uk - mark your query for the attention of the Thesaurus Team

More information about HASSET and ELSST

- Website:
  ukdataservice.ac.uk/about-us/projects/cessda-elsst/details.aspx

- Blog: https://elsst.wordpress.com/

- Announcements via HASSET Jiscmail list:
  HASSET-THESAURUS@Jiscmail.ac.uk

- Enquiries: help@ukdataservice.ac.uk – mark your query for the attention of the Thesaurus Team

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

UK Data Service

Questions

Contact details:

Lorna Balkan

balka@essex.ac.uk