From monolingual to multilingual thesaurus: the HASSET/ELSST relationship reviewed

Lorna Balkan
CESSDA Thesaurus Coordination Officer
UK Data Archive
University of Essex

NKOS, London

11-12 September 2014

UK Data Service

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

Overview

- Background: thesaurus development at UK Data Archive
- CESSDA-ELSST project
- challenges
- solutions – and some open questions

Background

- UK Data Archive holds the largest collection of digital research data in the social sciences and humanities in the UK
- long history of thesaurus development
- currently manages two related social science thesauri:
  - ELSST
  - HASSET

HASSET and ELSST

HASSET (Humanities and Social Science Electronic Thesaurus)

- over 40 years old – derived initially from the UNESCO thesaurus
- in-house thesaurus used for indexing and searching for data collections from the UK Data Service

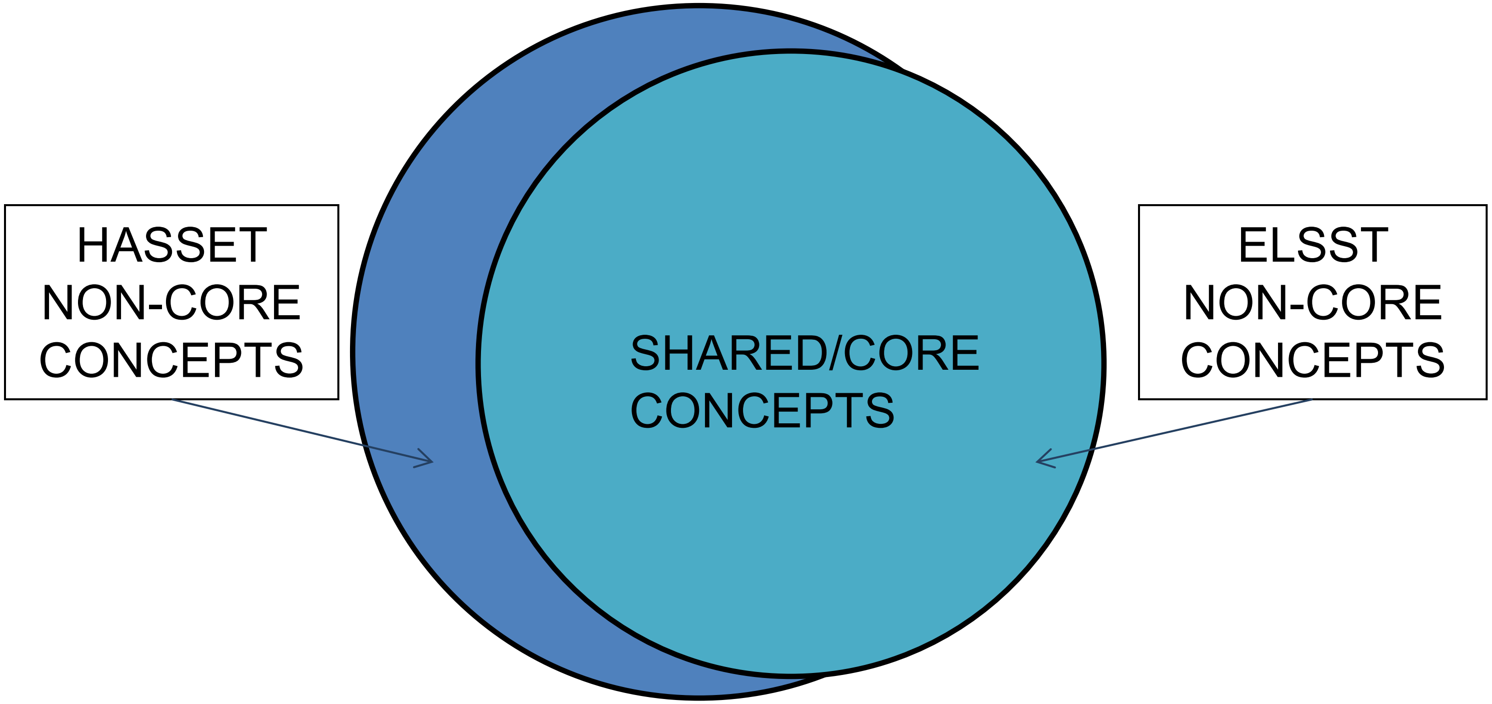
ELSST (Multilingual European Language Social Science Thesaurus)

- developed since 2000
- currently available in 9 languages, with more planned
- used for searching in CESSDA data portal
- derived from HASSET (English is source language)

HASSET and ELSST 2

- most ELSST concepts also in HASSET – “core concepts”
- both thesauri also contain non-core concepts
- 3286 concepts in ELSST
- 4743 concepts in HASSET
- all ELSST concepts must be internationally applicable
- non-core concepts in HASSET include
  - UK-specific (e.g. BARRISTERS)
  - other non-core concepts and hierarchies (e.g. GEOGRAPHICAL AREAS AND COUNTRIES hierarchy)

HASSET/ELSST relationship


HASSET NON-CORE CONCEPTS

SHARED/CORE CONCEPTS

ELSST NON-CORE CONCEPTS

HASSET and ELSST 3

- historically maintained on different platforms, but:
  - resource-intensive and inefficient
  - error-prone: thesauri may diverge without this being obvious or intentional
- HASSET and ELSST grown apart over the years

CESSDA-ELSST project

- 5-year ESRC-funded project (2012-2017)
- opportunity to revisit relationship between two thesauri
- initial hypothesis is that two thesauri could be merged
- Question: Is it possible/desirable?

CESSDA-ELSST project 2

Other project goals:

- update and revise both thesauri:
    - move from term-based to concept-based system
    - make ISO 25964-1 compliant (as much as possible)
    - create SKOS versions of both thesauri
- design new thesaurus management system
- streamline management processes

Can thesauri be merged?: Methodology

- define possible elements and relationships in each thesaurus
- define axioms and constraints *within* each thesaurus
- define axioms and constraints *between* two thesauri
- carry out alignment exercise
- refine axioms and constraints between two thesauri

Thesaurus alignment exercise

- aim: to see if feasible to merge ELSST and HASSET, i.e. to consider ELSST a true subset of HASSET
- method: looked at all concepts, terms and relations that were in ELSST, not HASSET and resolved where possible
- work to look at what is in HASSET, not ELSST, still to be done

Thesaurus alignment exercise 2

Conclusion:

- not possible/desirable to merge the two thesauri completely – relationship should instead be seen in terms of a mapping
- however, aim to make core concepts identical wherever possible, but allow divergence under certain circumstances
- aim to make the relationship between ELSST and HASSET concepts clear
- allows each thesaurus to retain own identity and integrity

ELSST → HASSET mapping relationship (core concepts)

- 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
- in both cases, other associated metadata may differ, including: UFs, NTs, RTs

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

UK Data Service

ELSST → HASSET equivalence versus ISO 25964-2 equivalence

ELSST/HASSET	ISO 25964-2
exact equivalence	exact simple equivalence
close equivalence	exact simple equivalence OR inexact simple equivalence

Types of mapping in ISO 25964-2

- basic mapping types:
  - Equivalence (denoted ‘EQ’)
    - Simple*: 1:1
    - Compound: 1: many
  - Hierarchical
    - Broader (denoted ‘BM’)
    - Narrower (denoted ‘NM’)
  - Associative (denoted ‘RM’)
- *exact vs. inexact: Inexact - concepts may:
  - be equivalent in some contexts but not others
  - have overlapping scopes or small differences of connotation

ELSST → HASSET equivalence versus ISO 25964-2 equivalence: 2

- ISO 25964-2 mappings are semantic
- ELSST → HASSET mapping:
  - has structural constraints (must have same BTs)
  - also requires identity of preferred labels
  - narrower (and more restrictive) case of the simple equivalence as defined by 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
- it is important to keep track of differences between core concepts in the two thesauri for thesaurus management purposes
- not all relationships between core concepts can be captured by axioms and constraints – reporting functions important

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:
  - SNs
  - UFs
  - NTs
  - RTs
- How will mapping be represented/implemented?

Different scope notes

- scope note in ELSST and HASSET defines or clarifies the semantic boundaries of a concept as it is used in the thesaurus
- does **not** include information and guidance on how terms may be used for indexing (contained in separate ‘use note’)
- we aim to make scope notes identical in each thesaurus wherever possible
- But is this always possible/desirable?
- system allows for flexibility

ELSST concepts

- ELSST concepts aim to be culture-neutral, i.e. applicable to all or most ELSST member countries, and contain no UK-bias
- but most ELSST concepts belong to social science domain
- many social science concepts have some element of culture-specificity – i.e. source culture may contain some elements and phenomena which do not exist or are different in the target culture

ELSST concepts: 2

Two types of culture-specificity:

- concept available cross-nationally, even if its meaning varies slightly from country to country

  Example: WELL-BEING

- concept not available at all in other language, and no term to describe (e.g. because of different systems of law, education, politics, etc.)

  Example: GRAMMAR SCHOOLS

- only first type of concept will be in ELSST

Different scope notes 2

- allowing HASSET and ELSST scope notes to vary would enable culture-specific (or country-specific) information to be added to the HASSET scope note, if required, while allowing ELSST scope notes to remain ‘neutral’
- difference in meaning would at most be difference between exact and inexact equivalence – no significant bearing on information retrieval
- scope notes will be reviewed as part of further alignment work

Different scope notes: putative example

- SOVEREIGNTY:

  “Supreme authority in a state. In any state sovereignty is vested in the institution, person, or body having the ultimate authority to impose law on everyone else in the state and the power to alter any pre-existing law. **In the UK Sovereignty is vested in Parliament**. In international law, it is an essential aspect of sovereignty that all states should have supreme control over their international affairs, subject to the recognized limitations imposed by international law.”

  (OXFORD DICTIONARY OF LAW)

Different UFs

- HASSET and ELSST UFs are allowed to differ
- E.g. PT in HASSET may be UF in ELSST:

HASSET	ELSST
SINGLE-SEX SCHOOLS NT: BOYS' SCHOOLS NT: GIRLS' SCHOOLS	SINGLE-SEX SCHOOLS UF: BOYS' SCHOOLS UF: GIRLS' SCHOOLS

- What about other UFs?

UFs in ELSST

- ELSST includes internationally relevant concepts only
- but country-specific UFs are allowed – useful as search aids
- no requirement to translate English source UFs
- same language version may have UF of different dialect (e.g. German version has Austrian UFs) – but these are not formally distinguished
- alternatively, different dialects of a language can have separate language versions (e.g. Mexican-Spanish planned in addition to Spanish-Spanish)

Different UFs: Options

- **either** delete UK-specific UFs from ELSST, and leave in HASSET only
- **or** allow some UK-specific UFs in ELSST, but limit their number and specify when allowable
- UK-specific UFs can be useful if preferred term very abstract
  - E.g. UPPER SECONDARY EDUCATION

    UF = SIXTH FORM EDUCATION
- decision likely to be on case-by-case basis

How will ELSST-HASSET mapping be represented/implemented?

- core concept mapping is currently for management purposes only - to keep track of differences between thesauri
- could also be useful for users, either visibly, or behind the scenes, to broaden search
- other mapping types could also be implemented later

E.g. ELSST core → HASSET non-core concept

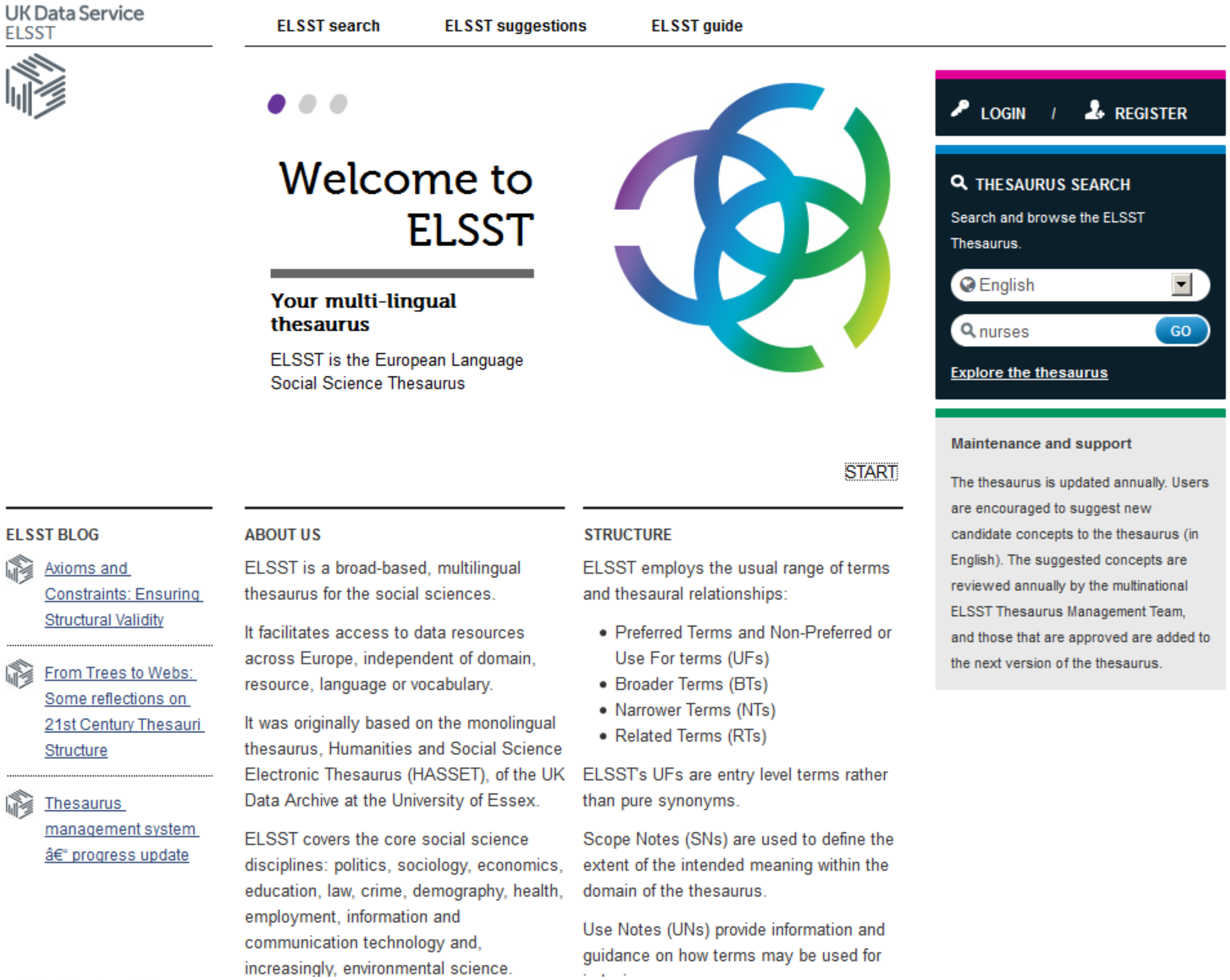
LOCAL TAXATION → COUNCIL TAX

(equivalent to LOCAL TAXATION **NM** COUNCIL TAX in ISO 25964-2)

Thesaurus management system (TMS)

- thesauri will not be merged, but mounted on single management system
- separate user-facing pages for both thesauri
- improved management interface
    - enables sharing of suggestions with colleagues inside and outside the Archive
    - keeps track of changes – terms never deleted
    - keeps track of differences between two thesauri

TMS user-facing interface: ELSST homepage



UK Data Service
ELSST

ELSST search | ELSST suggestions | ELSST guide

● ○ ○

Welcome to ELSST

Your multi-lingual thesaurus

ELSST is the European Language Social Science Thesaurus

START

LOGIN / REGISTER

THESAURUS SEARCH

Search and browse the ELSST Thesaurus.

English

nurses GO

Explore the thesaurus

Maintenance and support

The thesaurus is updated annually. Users are encouraged to suggest new candidate concepts to the thesaurus (in English). The suggested concepts are reviewed annually by the multinational ELSST Thesaurus Management Team, and those that are approved are added to the next version of the thesaurus.

ELSST BLOG

- Axioms and Constraints: Ensuring Structural Validity
- From Trees to Webs: Some reflections on 21st Century Thesauri Structure
- Thesaurus management system â€" progress update

ABOUT US

ELSST is a broad-based, multilingual thesaurus for the social sciences.

It facilitates access to data resources across Europe, independent of domain, resource, language or vocabulary.

It was originally based on the monolingual thesaurus, Humanities and Social Science Electronic Thesaurus (HASSET), of the UK Data Archive at the University of Essex.

ELSST covers the core social science disciplines: politics, sociology, economics, education, law, crime, demography, health, employment, information and communication technology and, increasingly, environmental science.

STRUCTURE

ELSST employs the usual range of terms and thesaural relationships:

- Preferred Terms and Non-Preferred or Use For terms (UFs)
- Broader Terms (BTs)
- Narrower Terms (NTs)
- Related Terms (RTs)

ELSST's UFs are entry level terms rather than pure synonyms.

Scope Notes (SNs) are used to define the extent of the intended meaning within the domain of the thesaurus.

Use Notes (UNs) provide information and guidance on how terms may be used for

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

UK Data Service

Summary and conclusion

- thesaurus alignment task made it possible to review relationship between HASSET and ELSST
- HASSET and ELSST will not be merged, but mapped
- core concepts will be identical wherever possible, but divergence will be allowed
- allows thesauri to preserve own identity and integrity
- some inter-thesaural relationships can be expressed via axioms and constraints – others left for manual control
- reporting functions in new thesaurus management system will allow developers to keep track of differences between two thesauri

Obtaining HASSET and ELSST

- both free to browse:
  - HASSET: http://hasset.ukdataservice.ac.uk
  - ELSST: http://elsst.ukdataservice.ac.uk
    (Both in beta at present)

- 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

Questions

Contact details:

Lorna Balkan

balka@essex.ac.uk