Innovation in Thesaurus Management

Lucy Bell
Management Information Manager
UK Data Archive

IASSIST 2013, Cologne
31 May 2013
Two thesauri; two projects

- **SKOS-HASSET**
  - 10 month, Jisc-funded project to enhance HASSET

- **CESSDA ELSST**
  - 5-year, ESRC-funded project (first 18 months for development) to re-develop the thesaurus management applications and streamline their processes

- …together, both projects should create efficiencies
The UK Data Service: cataloguing standards

- The UK Data Service indexes over 5000 digital data collections – and the number is ever growing
  - All catalogued at thematic level
  - Many also indexed at variable level
- Available via Discover (discover.ukdataservice.ac.uk):
  - HASSET is employed in the search
  - And every data collection is indexed with HASSET terms
Cataloguing practices

• professional, human indexers
  • review study documentation
  • review data files
  • apply HASSET keywords that represent the data

• consistency: attention paid to terms used over time within data series and across similar studies

• time: varies considerably depending on the size and complexity of the study, but resource-intensive

• other things that are indexed:
  • questionnaires; case studies; support/how to guides
HASSET

- multidisciplinary thesaurus developed originally to support the UK Data Archive/UK Data Service collections
- coverage in the core subject areas of social science
- uses standard hierarchical relationships:
  - TT (top term)
  - BT (broader term)
  - NT (narrower term)
  - RT (related term)
  - USE (from non-preferred term to preferred term)
  - UF (from preferred term to non-preferred term)
- constantly updated
- role of HASSET in the Archive is twofold:
  - used internally for **indexing** studies and series with HASSET terms
  - also a separate **product** licensed to others
ELSST

• European Language Social Science Thesaurus (ELSSST) is a multi-lingual thesaurus, based on core English terms taken from HASSET
• mirrors HASSET’s structure
• translated into 8 languages (with more on the way)
• closely connected with HASSET, but must demonstrate international applicability of all terms
• permits longer lead-in time for concept consensus
• moving to annual release
SKOS-HASSET project

• high level aim: to enhance both aspects of HASSET
  • (its use internally for indexing
  • its use externally as a licensed product)

• three stated objectives:
  • apply SKOS to HASSET
  • improve its online presence
  • test its automated indexing capabilities
What is SKOS?

- **SKOS/RDF** – what is RDF?
  - RDF describes a piece of data using a simple format of subject – predicate – object
    - E.g., *car* – hascolour – *red*

- **So, what is SKOS?**
  - Simple Knowledge Organization System
  - SKOS is set of RDF predicates to describe relationships between thesaurus terms
    - E.g., `skos:concept162` – `skos:preflabel` – “CAR”
    - E.g., `skos:concept162` – `skos:altlabel` – “AUTOMOBILE”
  - it encodes these products in a standardised way to make their structures comparable and to facilitate interaction
Applying SKOS to HASSET

- SKOS has been applied to HASSET: [lod.data-archive.ac.uk/skoshasset/](lod.data-archive.ac.uk/skoshasset/)
- persistence via GUIDs (Globally Unique IDentifiers)
- version control
- we are using Pubby to publish our SKOS
- and BrightStarDB for the Triple Store
SKOS to HASSET: issues

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

• Licence Recommendation Report, issued via our blog
• big question: which, if any, licence to use?
• releasing HASSET with no contract might compromise its integrity
• we reviewed 12 online thesauri:
  • 8 (66%) used licences:
    • 5 (42%) had Creative Commons licences
    • 3 (25%) maintained their own, bespoke licensing arrangements
  • 4 (33%) were freely available
• the licensing landscape contains many, varied features, with different licence models in operation
Licensing risk analysis

‘the priorities of the initial licensor of the content need to be based upon an open vs risk evaluation, rather than openness only’

Korn, N. (2011) Overview of the ‘Openness’ of licences to provide access to materials, data, databases and media. JISC/Naomi Korn Copyright Consultancy, January 2011.
Licence recommendation: a bespoke licence

- can be adapted for both HASSET and ELSST and any combinations of their shared terms (efficiency here!)
- has the flexibility to allow all users to gain access to the product
- does not permit the thesaurus being passed on to third parties
- can be applied to a multi-IP situation
Automated indexing: four corpora (33,959 files)

- Nesstar questions/variables (humanly indexed during project) - 26,753 files
- Questionnaires - 1,353 files
- Catalogue records - 5,610 files
- Publications (case studies / support/how to guides) - 243 files

Case study - research

Entering the NEET zone: Does career indecision matter?

Author: Ricardo Sabates, University of Sussex, Scott Yates of De Montfort University, Angel Harris of Princeton University and Jeremy Staff of Pennsylvania State University

Date: 4 December 2012

Type of case study: Research

About the research

A lot of government and media attention has recently focused on young people who are classified as 'NEET' (not in employment, education or training). This in turn has created widespread interest in examining the underlying factors that tend to disrupt career and educational paths early in adult life.

This study draws on the British Cohort Study to examine two possible factors: early uncertainty and misalignment in occupational aspirations. The findings indicate that young people with uncertain or misaligned aspirations at age 16 are much more likely to become NEET by age 18. Furthermore, uncertainty and misalignment factors are both more widespread and more detrimental for those from poorer backgrounds.

These research findings suggest that, relative to young people with high and aligned ambitions and those with higher educational expectations than their aspired career, young men with mismatched aspirations and educational expectations were almost twice as likely to become NEET. Young women with misaligned ambitions were three times as likely to become NEET.

In short, this study illustrates that there is certainly a link between young people’s aspirations for future occupations and their labour market outcomes. Educational attainment appears to be increasingly important in shaping young people’s life chances and transition to adulthood – something likely to have impacted especially the poorest and lowest-achieving young people. Finally, realistic occupational expectations depending on educational performance and adequate knowledge of available post-16 job...
Keyphrase Extraction Algorithm (KEA)

- created a training model using human indexer’s keywords
- 80% of text used for training / 20% of text used for testing
- with SKOS-HASSET as controlled vocabulary
Automated indexing: results

- little overlap between KEA keywords and manual keywords
- however, a high percentage of KEA keywords were considered relevant/suitable even if they were not exact matches:
  - 33% for the SQB corpus
  - with an average of 25% across all four corpora
- **KEA could be a very useful tool for indexers**

- See poster session!
Other things we did

- online user web pages refreshed and user guidance provided: [www.data-archive.ac.uk/find/hasset-thesaurus/skos-hasset](http://www.data-archive.ac.uk/find/hasset-thesaurus/skos-hasset)

- project web site: [www.data-archive.ac.uk/find/our-projects/skos-hasset](http://www.data-archive.ac.uk/find/our-projects/skos-hasset)

- SKOS-HASSET blog: [hassetukda.wordpress.com/](http://hassetukda.wordpress.com/)

- Laid the groundwork for the CESSDA ELSST project…
CESSDA ELSST project

- 5-year ESRC-funded project (2012 – 2017)
- first 18 months = technical development time
- aim: root and branch re-development of the ELSST and HASSET interfaces, structures and processes
  - merging the two existing management UIs internally
  - externally updating and improving the user-facing pages
  - reviewing the thesauri’s structures
CESSDA ELSST project: deliverables

- improved management interface
- streamlined access to the thesauri/improved user experience
- ideas... from this...
- to this?
- or another form based system?
- or something else?
streamlined management process, internally
  • single workflow for concept management (alignment work)
  • Finite State Machine
  • fulfilling two different time-based requirements
implement annual release process (with international concept consensus)
improved and ISO-compliant (as much as possible!) thesauri (ISO 25964-I and 25964-II)
SKOS-ELSSST
CESSDA ELSST: what have we done so far?

- project management/preparation:
  - project documentation / requirements gathering:
    - consultation with CESSDA colleagues
    - questionnaire for ELSST translators
  - review of ISO 25964-1 implications
  - review of interoperability issues in relation to HASSET and ELSST
- translations (with Taina):
  - translation guidelines for ELSST reviewed, updated and re-released
  - translation meeting procedures for ELSST created and trialled
  - online training given to new Italian, Czech and Romanian ELSST translators
- access/distribution:
  - work begun on new distribution agreement
  - procedures reviewed
CESSDA ELSST: what have we done so far? (2)

- Information development - consolidation
  - Subject categories <topcClas> mapped to HASSET/ELSST concepts
  - Feasibility of consolidated structure considered
    - Currently, excluding geographies, 297 TTs
    - Considering more traditional, more easily browseable tree structure (cf. MeSH with 16 TTs)
CESSDA ELSST: concept alignment

- Inconsistencies between HASSET and ELSST identified and work begun to align the two ...
- 7,695 concepts
- 4,032 synonyms

- 101,808 triples (relationships)
  - 48,430 in HASSET (24,439 in HASSET only)
  - 53,378 in ELSST (28,782 translations)

- 23,991 triples shared by HASSET and ELSST

- 605 triples (representing 242 concepts) in ELSST only – being worked on right now
CESSDA ELSST project: communication

- investigating methods of communication and collaboration with our translation partners:
  - wiki?
  - virtual meeting/email communication with translators expected soon
  - blog: elsst.wordpress.com
  - announcements via the HASSET Jiscmail list: HASSET-THESAURUS@Jiscmail.ac.uk

- any questions may be directed to the team: help@ukdataservice.ac.uk, marking your query for the attention of the Thesaurus Team
Questions

Contact details:
Lucy Bell
lajbell@essex.ac.uk