

---

# Enhancing metadata quality

Lucy Bell, Anne Etheridge

UK Data Archive, University of Essex

CES14, UK Data Archive, University of  
Essex

14 November 2014

---

UK Data Service

---



---

# Metadata

- Metadata retrieval, with high precision, is integral to the Service's aim of providing its users with seamless access to a wide range of data collections
- Resource discovery mechanisms and tools have been given high priority
- The vision: a world of accessible metadata, connecting resources, presenting the user with a map of possible pathways through the data



---

# What constitutes quality metadata?

- Schema?
- Openness?
- Shareability?
- Versioning?
- Currency?
- Comprehensiveness?
- Connectedness/whole system view?
  
- All of the above?
- Balance with local needs?
  
- How can we tell? What are the impact criteria?  
(Metrics/logs?)



---

# Open data vs open metadata

- Open metadata
  - Discovery Open Metadata Principles:
    - “Open metadata creates the opportunity for enhancing impact through the release of descriptive data about library, archival and museum resources. It allows such data to be made freely available and innovatively reused to serve researchers, teachers, students, service providers and the wider community in the UK and internationally.” (Discovery, 2012)
  - Metadata are, by default, made freely available for use and reuse unless explicitly precluded by third party rights or licences



# Connecting metadata

- Key = links
- Links made between information resources (e.g. data collections, how to guides, related outputs, case studies)
- Metadata encoded so that they may be linked with and to external resources as well, especially cross-Europe
- Wider initiatives in other fields, see:
  - Natural Europe project ([http://ec.europa.eu/information\\_society/apps/projects/factsheet/index.cfm?project\\_ref=250579](http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=250579))
  - Functional Requirements for Bibliographic Records (FRBR)



---

# Implementation of the vision

- Metadata schema used by the UK Data Service include:
  - Data Documentation Initiative Codebook (DDI-C 2.5) from the DDI Alliance
  - QuDex
  - Text Encoding Initiative (TEI)
  - SDMX
  - Dublin Core (DC)
- Schema are also mapped to MARC 21, METS and MODS and UK GEMINI2 to ensure the greatest compatibility possible



# DDI

- DDI-Lifecycle (DDI-L)
  - Encompasses DDI-Codebook specification and extends it
  - Designed to document and manage data across the entire life cycle, from conceptualization to data publication and analysis and beyond
  - Based on XML Schemas
  - Modular and extensible
- DDI-Codebook (DDI-C)
  - more light-weight version of the standard
  - intended primarily to document simple survey data
  - Originally DTD-based
  - now available as an XML Schema
  - More similar to a traditional, bibliographic schema
- UK Data Service uses DDI-Codebook



---

# Discover

- Metadata connections are visible via Discover, the Service's search and browse application
- Uses SOLR technology
- In the spirit of FRBR, Discover promotes connections between:
  - related data
  - case studies of data use
  - publications and other outputs
  - citations
  - series



# Discover

## Discover

- Discover
- Variable and question bank
- QualiBank
- Type
  - Data collections (3099)
  - Case studies (77)
  - Series (48)
  - ESRC outputs (321)
  - Support / how to guides (22)
- Subject +
- Date +
- Data type +
- Key data +
- Country +
- Kind of data +
- Spatial unit +
- Analysis unit +
- Access +
- Depositor +
- Teaching data +

Login may be disrupted between 07.00 and 09.00 on Tuesday 23 September 2014 during system upgrades. Secure Lab access will not be affected.

Search and browse our data collections, support guides, case studies, and related publications.

health

[Reset filters](#) | [Clear search](#) |  Auto-complete |  Map search to HASSET thesaurus? | [Help](#)

- Case study
- Data collection
- Series record
- ESRC output
- Support guide

Sorted by:

SEARCH RESULTS

Displaying 1-10 of 3099 results

1 2 3 4 5 >>>

- SN 7301 ONS Opinions Survey, Well-Being Module, March 2010  
Office for National Statistics. Social Survey Division  
[Full record...](#)  
[nesstar Explore online](#) | [Download/Order](#) | [Get full DDI XML](#) | [Similar data collections](#)
- SN 5232 Young People's Situations and Well-being in Siberia, 2002-2003  
Glendinning, A., University of Aberdeen. Department of Sociology  
[Full record...](#)  
[Download/Order](#) | [Get full DDI XML](#) | [Similar data collections](#)
- SN 6994 Annual Population Survey: Subjective Well-Being, April - September, 2011  
Office for National Statistics. Social Survey Division  
[Full record...](#)  
[Download/Order](#) | [Get full DDI XML](#) | [Similar data collections](#)
- SN 850712 The unanticipated transformative potential of online forums: Implications for well-being and offline social impact  
Louise Pendry, University of Exeter

Discover links data to case studies, series, outputs, how to guides, similar data collections and publications

---

# Persistent identification and data citation

- DataCite Digital Object Identifiers (DOIs) added to all data collections
  - Discoverable through DataCite Metadata Search
- ODIN
  - UK Data Service is a stakeholder in the BL's ODIN project, looking at linking DOIs with ORCIDs
- Granular data citation methodologies
  - Paragraph-level citation within QualiBank
  - Dynamic: allows a user to indicate which user-selected parts of a text have been selected as the material to reference
  - Generates on the fly references, ready to be copied and pasted into any reference list



# Paragraph-level citation

The image displays a sequence of overlapping screenshots from the UK Data Service website, illustrating the process of generating a paragraph-level citation. The screenshots are arranged in a cascading manner, showing the user's progression through the site's interface.

**Top Screenshot (Main Site):** Shows the UK Data Service homepage with navigation links: About us, Get data, Use data, Manage data, Deposit data, and News and Events. The breadcrumb trail is Discover > QualiBank > Document. The document title is "Interview with Mrs. Rayner".

**Second Screenshot (Discover Page):** Shows the "Discover" page for "Interview with Mrs. Rayner". It includes a sidebar with "Variable and question bank" and "QualiBank" selected. The main content area shows the document title and navigation options.

**Third Screenshot (Details Page):** Shows the "DETAILS" page for the document. It includes a sidebar with "External resources" and "QualiBank" selected. The main content area shows the document title and navigation options.

**Fourth Screenshot (Citation Generation Tool):** Shows the "Citation" tool. It includes a sidebar with "Create citation" and "Select an extract" selected. The main content area shows the document title and navigation options.

**Fifth Screenshot (Citation Result):** Shows the "Citation" result. It includes a sidebar with "Cancel" and "4 extracts selected" selected. The main content area shows the citation text and instructions.

**Citation Text:** A unique citation reference has been generated based on your selection.

Thompson, P., University of Essex. Department of Sociology, Lummis, T., University of Essex. Department of Sociology: "Interview with Mrs. Rayner" in "Family Life and Work Experience Before 1918, 1870-1973" 7, UK Data Service [distributor], 2009-05-12, SN:2000, Paragraphs 1-4. <http://dx.doi.org/10.5255/UKDA-SN-2000-1>, <http://discover.ukdataservice.ac.uk:/QualiBank/Document/?cid=q-943fb156-212a-4ee1-96eb-8e84e9c20870>

**Instructions:**

- Select the text above**
- You can copy and paste this citation as required in your outputs. This citation includes a URL which will link directly back to this page, where the cited text will be highlighted.
- Show preview of citation URL in action**

---

# Citation issues

- Citing at variable level?
- Citing dynamically updated data?
- Encouraging researchers to cite?



---

# CVs: the application of quality

- UK Data Service's roles:
  - Creation
  - Use
- DDI Controlled Vocabularies Group
  - Sets up CVs to populate elements in DDI-L and DDI-C
- Service manages two well-used thesauri:
  - HASSET (Humanities and Social Science Electronic Thesaurus)
  - ELSST (European Language Social Science Thesaurus)



---

# ELSST

- Traditional, hierarchical structure
- Uses TTs, BTs, NTs, RTs, UFs (as entry level terms) and Scope Notes
- SKOS version coming
- Recent ESRC award, CESSDA ELSST, has provided the funds to re-develop the management application:  
<http://elsst.ukdataservice.ac.uk>
- It is undergoing user testing
- CESSDA ELSST has also provided funds to review, refine and improve the thesaurus's structure



# ELSST application

UK Data Service  
ELSST

ELSST search   ELSST suggestions   ELSST guide

Elst > Thesaurus search

### Thesaurus search

Search and browse functionality

UK Data Service  
ELSST

ELSST search   ELSST suggestions   ELSST guide

Elst > Thesaurus search > View term

### View term

View full concept metadata and all translations

UK Data Service  
ELSST

ELSST search   ELSST suggestions   ELSST guide

Elst > Thesaurus search > View term > Visual graph

### View term

Home   New search   Login

Search current

Enter preferred language

HEALTH PROFESSIONALS

Visual graph

Reset Graph   Fullscreen

UF = Use For   BT = Broader Terms   NT = Narrower Terms   RT = Related Terms   = Expanded Term

Term   Click any term to expand and view related terms   Click icon to navigate to term details page

DISTRICT NURSES

NURSES

MIDWIVES

HEALTH PROFESSIONALS

Visual graph showing relationships between terms: DISTRICT NURSES (UF), NURSES (RT), HEALTH PROFESSIONALS (BT), and MIDWIVES (NT).

Tree view



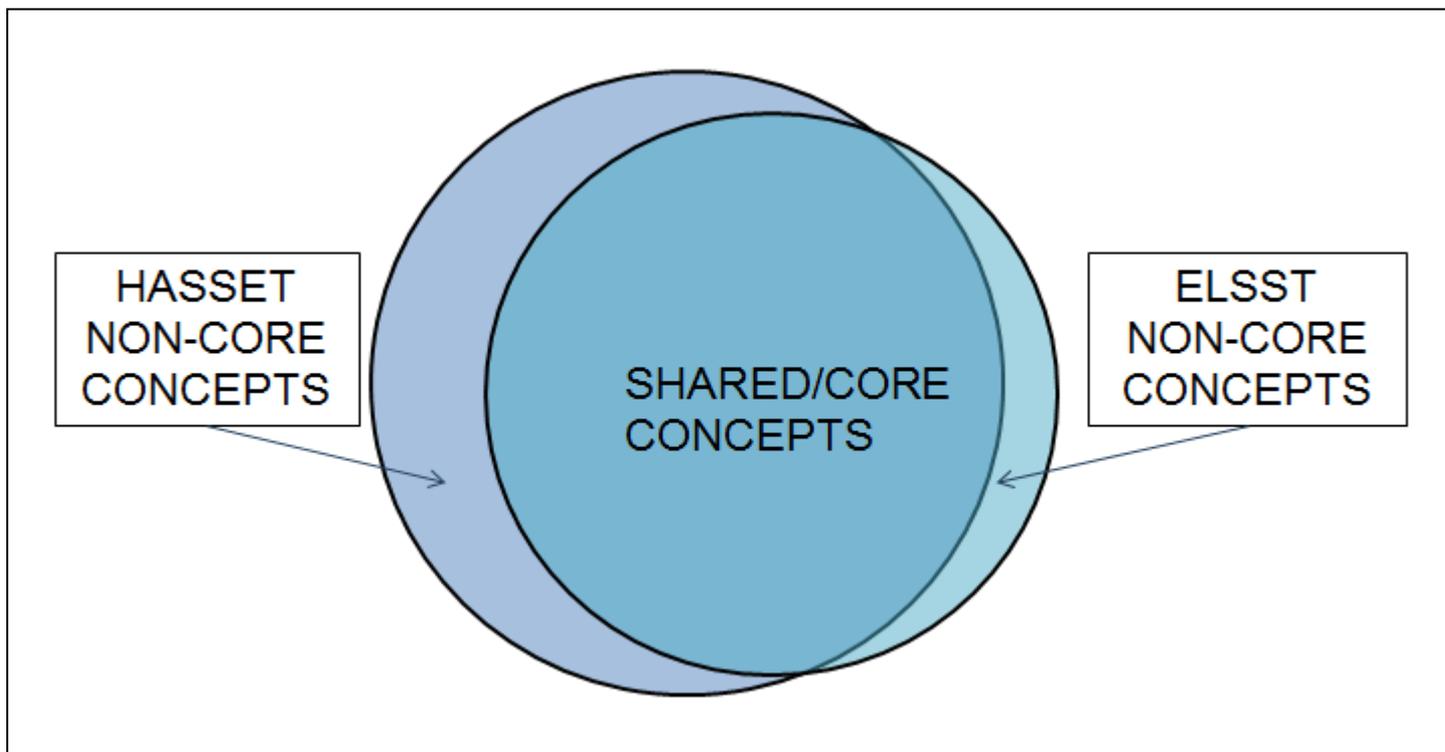
---

# ELSST

- Originally developed as part of the EU-funded LIMBER project in 2002. Further enhanced through additional funding from the ESRC and the University of Essex and subsequent EU grants (e.g. MADIERA and PPP)
- Higher-level concepts, with international applicability
- Exists in ten languages, with more on the way
- Team of global translators who provide advice and guidance on its development



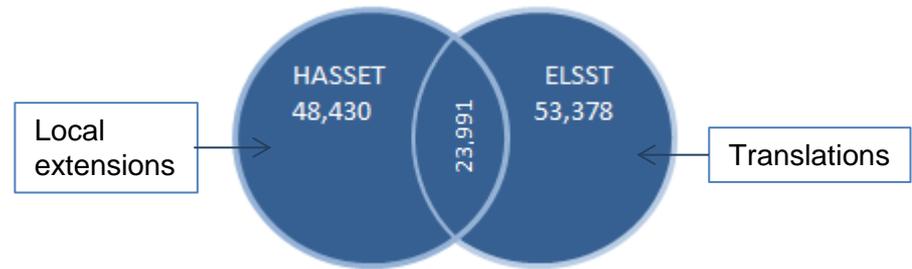
# CESSDA ELSST: concept alignment



- Historically, HASSET and ELSST have been managed separately, even though they share core concepts and structures
  - Time-consuming and not error-proof
  - inconsistencies between HASSET and ELSST identified and changed

# CESSDA ELSST: concept alignment

- 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

---

# Creating a shared set of concepts

- Based on, but extending, principles from ISO 25964 (NISO 2011, 2013)
- Concepts mapped using extended equivalence relationships
- Non-core concepts considered different in each thesaurus
- Core concepts have some degree of equivalence:
- Exact equivalence achieved with the same:
  - Preferred Term (PT) – giving linguistic integrity
  - Broader Term(s) (BTs) – giving structural integrity
  - Scope notes – giving semantic integrity
  - Scope note sources
- Close equivalence achieved with the same:
  - Preferred Term (PT)
  - Broader Term(s) (BTs)



---

# Creating a shared set of concepts

- Ensures divergence is accommodated
- Some concepts are very close cross-nationally, but require slightly different Scope Notes for British English (in HASSET), e.g.
- **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)



---

# Creating a shared set of concepts

- Another example: SOCIAL ASSISTANCE
- The ELSST definition refers to insurance:

“Assistance in money or in kind to persons, **often not covered by social insurance**, who lack the necessary resources to cover basic needs.”

The HASSET definition does not include this reference, because social insurance does not exist in the UK:

“Assistance in money or kind to persons whose income is below a certain level and who lack the necessary resources to cover basic needs.”



# Structural improvements

- 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)

1. [+](#) Anatomy [A]
2. [+](#) Organisms [B]
3. [+](#) Diseases [C]
4. [+](#) Chemicals and Drugs [D]
5. [+](#) Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. [+](#) Psychiatry and Psychology [F]
7. [+](#) Phenomena and Processes [G]
8. [+](#) Disciplines and Occupations [H]
9. [+](#) Anthropology, Education, Sociology and Social Phenomena [I]
10. [+](#) Technology, Industry, Agriculture [J]
11. [+](#) Humanities [K]
12. [+](#) Information Science [L]
13. [+](#) Named Groups [M]
14. [+](#) Health Care [N]
15. [+](#) Publication Characteristics [V]
16. [+](#) Geographicals [Z]

MeSH

- [-] HASSET
  - [+](#) ABILITY
  - [+](#) ACHIEVEMENT
  - [+](#) ADMINISTRATION
  - [+](#) ADMINISTRATION OF JUSTICE
  - [+](#) ADMINISTRATIVE AREAS
  - [+](#) ADMINISTRATIVE STRUCTURES
  - [+](#) ADVICE
  - [+](#) AGE
  - [+](#) AGE GROUPS
  - [+](#) AGRICULTURE
  - [+](#) AGRONOMY
  - [+](#) AID
    - AMNESTY
  - [+](#) ANALYSIS
  - [+](#) ANIMAL HUSBANDRY
  - [+](#) ANIMALS
  - [+](#) ANTHROPOLOGY
  - [+](#) ARMAMENT PROCESS
  - [+](#) ARMED FORCES
  - [+](#) ARTS
    - ASTRONOMY
  - [+](#) ATTENDANCE
  - [+](#) ATTITUDES
    - BEHAVIOURAL SCIENCES
  - [+](#) BELIEFS
  - [+](#) BIOLOGICAL CONTROL
  - [+](#) BIOLOGY
  - ...



---

# Questions for discussion

- What are quality metadata?  
(is it gleaned via data usage in fact – if data can be found, then the metadata work?)
- How can we improve this further?  
(how do we know which data are not being found? What role do logs and metrics play?)
- How to encourage data citation?
- How to cite at a more granular level?
- How to cite dynamically updated/streamed data?
- How can we make better use of CVs and thesauri?

