SEEMINGLY GLIDING: THE POWER OF METADATA IN ACADEMIC RESOURCE DISCOVERY SYSTEMS

LUCY BELL

MANAGEMENT INFORMATION MANAGER
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

ANAT VERNITSKI

BIBLIOGRAPHIC SERVICES OFFICER
UK DATA ARCHIVE
UNIVERSITY OF ESSEX

UK•DATA ARCHIVE

FACETS OF KNOWLEDGE ORGANIZATION, ISKO 2011, **UCL**

04 JULY 2011

Introduction

- trends in information-seeking behaviour
- example of a powerful metadata tool: HASSET
    - indexing
    - HASSET's development
    - extension into a multi-lingual tool
- review of the UK Data Archive's resource discovery tools
- a vision for resource discovery
- applying the vision, using the existing tools

UK **DATA ARCHIVE**

How do researchers find data in 2011?

- trends in information-seeking behaviour show that users prefer simple, Google-like interfaces…

- …but which still return acutely-focused and highly-relevant results.

- the look and feel of the interfaces should be simple but the results must achieve academic rigour.

UK **DATA ARCHIVE**

The metadata conundrum

- for data services to produce simple interfaces - which still return highly-relevant results - metadata are required which are both:
  - extremely powerful
  - increasingly invisible

- a conceptual shift has taken place: the work to focus searches has moved behind the interface

- a challenge exists for services to create resource discovery interfaces which are supported by **high-functioning, yet hidden, metadata**.

UK **DATA ARCHIVE**

HASSET as resource discovery tool within UK Data Archive

- Humanities and Social Sciences Electronic Thesaurus (HASSET);

- in-house thesaurus;

- developed by the Archive over the past 30 years or so, initially based on the UNESCO Thesaurus (Jean Aitchison,1970s);

- over time HASSET has been continuously expanded and updated for use as an online retrieval system;

HASSET as resource discovery tool within UK Data Archive

- HASSET is a multidisciplinary thesaurus developed initially to support the UK Data Archive collection;

- coverage more comprehensive in the core subject areas of social science disciplines, including politics, sociology, economics, education, law, crime, demography, health, employment, and, increasingly, technology and environmental studies;

- humanities disciplines such as history and linguistics also have a prominent presence;

- the role of HASSET in the Archive is twofold:

  - it is used as a tool for adding value and improving resource discovery by indexing particular studies and series with appropriate HASSET keywords;

  - it is also a separate product developed in-house in the Archive and, as such, reflects the expertise that the organization possesses.

HASSET page on the UK Data Archive

The HASSET search interface online

Search box leading to HASSET terms and hierarchies

UK **DATA ARCHIVE**

Indexing using HASSET

- HASSET is used by cataloguers in the Archive to index datasets compiled into studies;

- keyword indexing is done in conjunction with cataloguing;

- cataloguers receive in-house initial training as well as refresher training in cataloguing and indexing;

- regular meetings are held to share good practice and discuss updates and queries.

Indexing using HASSET

- quantitative studies indexing options:
    - a keyword is assigned to each variable;
    - a keyword is assigned to each question or group of questions in a survey;

- qualitative studies indexing options:
    - a keyword is assigned to each interview schedule topic, interview topic or any other topic occurring in qualitative types of data such as diaries, focus group notes or observation field notes.

The development of HASSET

- currently, new HASSET terms are developed regularly in order to cover the ever-growing collection of the UK Data Archive;

- cataloguers often flag up the need for new vocabulary to index a particular study;

- each request is analysed in the context of the collection as a whole;

- considering the place of a new term in the hierarchy is one of the main considerations when creating new terms;

The development of HASSET

- HASSET uses the following 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).

HASSET interface displaying hierarchies

Searching data indexed by HASSET terms

UK **DATA ARCHIVE**

HASSET search results

List of studies in the online catalogue indexed with the chosen HASSET keyword

HASSET as a source for multilingual thesauri

- HASSET is the source for the multilingual thesaurus European Language Social Science Thesaurus (ELSST), building on the projects LIMBER and MADIERA;

  - the LIMBER project, 2000-2001, developed multilingual tools to support user access to the data stored at social science archives across Europe and to integrate with data from other domains;

  - following on from LIMBER, the MADIERA project, 2002-2006, aimed to create an effective operational web-based infrastructure for the European social science community.

HASSET in use beyond the UK Data Archive

- the UK Data Archive was approached already in the early 1980s by outside organisations who were interested in using HASSET for their own purposes;

- in 1997 permission was received from UNESCO to consider HASSET as a separate product;

- since then, HASSET continues to be used by organizations outside the Archive.

HASSET in use beyond the UK Data Archive

- external users normally wish to receive a copy of HASSET for

  - evaluation
  - to be used in a project
  - or to become the basis for separate systems tailored to their needs.

- several prestigious organizations use HASSET as a controlled vocabulary to serve their own collections

- other organizations use HASSET to inform the development of their own controlled vocabularies.

Enhanced resource discovery in a service

- the UK Data Archive, based at the University of Essex, is the curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. It holds several thousand datasets relating to society, both historical and contemporary. It has been in existence for over 40 years.

- May – October 2010: a review was carried out of the UK Data Archive's resource discovery tools.

Reviewing the UK Data Archive's tools

- the primary resource discovery tools reviewed:
  - ESDS (www.esds.ac.uk)
  - UK Data Archive (www.data-archive.ac.uk)
  - Census Portal (www.census.ac.uk)
  - Survey Question Bank (www.surveynet.ac.uk)
  - History Data Service (hds.essex.ac.uk)
  - RELU-DSS (relu-data-archive.ac.uk)
  - controlled vocabularies (**HASSET and subject categories, supporting the search experience**)
  - plus a comparison of resource discovery elsewhere

- the brief:
  - review the Archive's resource discovery tools making recommendations on how they can be improved;
  - undertake a detailed examination of the website;
  - describe each method of resource discovery;
  - document all existing internal procedural documents, standards documents and current user guidance noting gaps;
  - collate evidence from existing reviews about user satisfaction and usability.

- White Paper was submitted to the Archive and ESDS SMTs in November 2010.

The wider Archive context: 21 interfaces

HASSET and other CVs may be used in the majority of search and browse activities.

UK **DATA ARCHIVE**
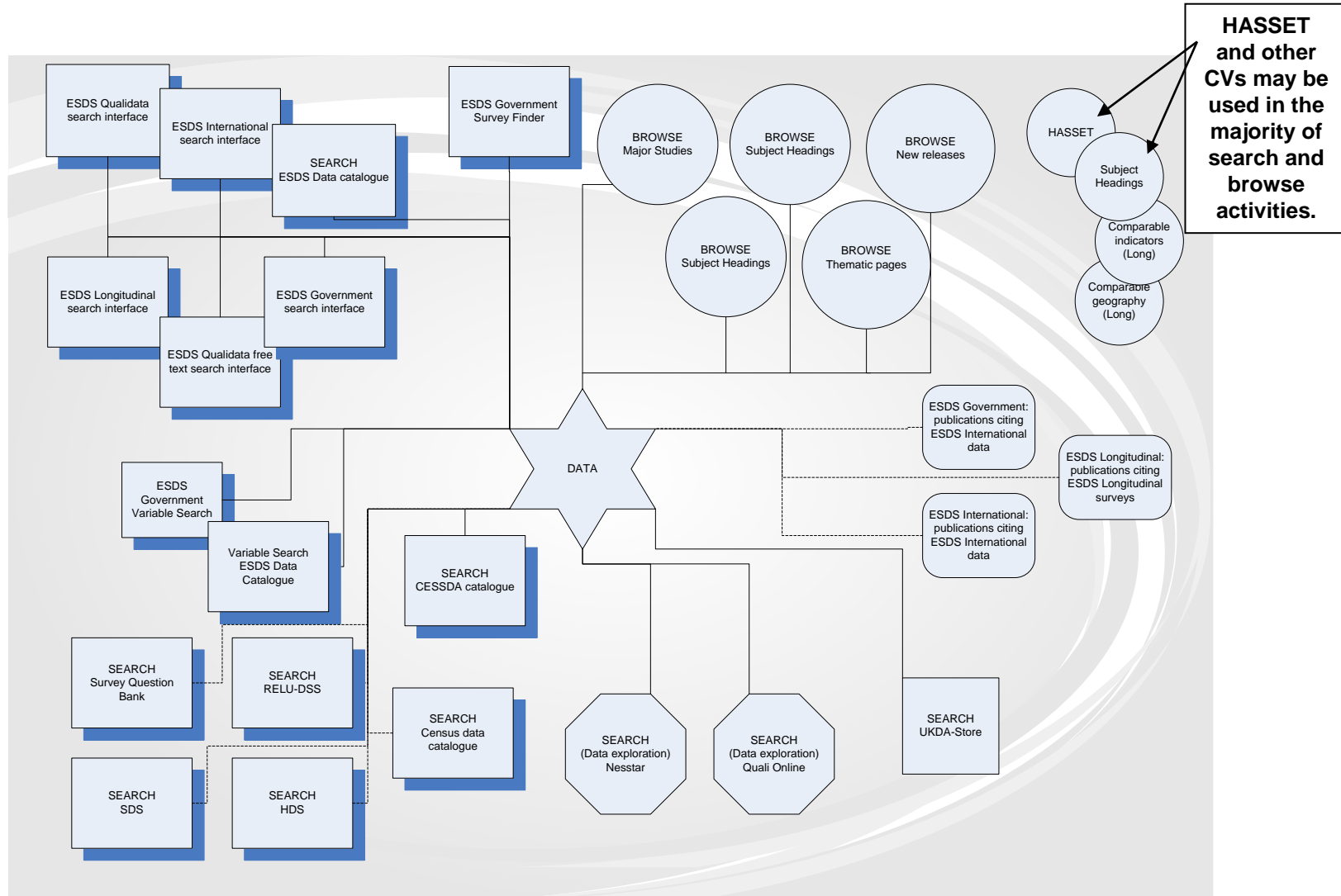
The issue:

- *how to integrate these interfaces and the metadata?*

The vision:

- *a single search interface,*
- *plus, at the results, the ability to move seamlessly from one service/type of data to another:*
  - *via faceted browsing and*
  - *directly from within each resource type*

- it means cross-referencing ESDS data collections with publications, with research outputs, with citations, with question text with web pages and beyond
  - using **metadata**
  - in the context of a move to **DDI 3**

Richer metadata and tools

- building on work which allows **cross-service searching** and makes **linkages** between data held in different services
  - from question text in SQB to data visualisation in Nesstar
  - from RELU-DSS research outputs to data
  - and maintaining our OAI presence

Richer metadata and tools

- controlled vocabularies:
  - HASSET: augmented role within the search interface
  - DDI 3 version
  - controlled vocabulary for outputs

- extend HASSET's uses:
  - gather more metadata from depositors
  - use metadata to make connections via
    - relationship analysis
    - just-in-time indexing of searches/results/user backgrounds to present more data that may be of interest

- other Automatic Metadata Generation techniques (such as harvesting from other websites (ESRC, ONS etc.); text mining/geographic coordinates)

Richer manual/automatic metadata generation
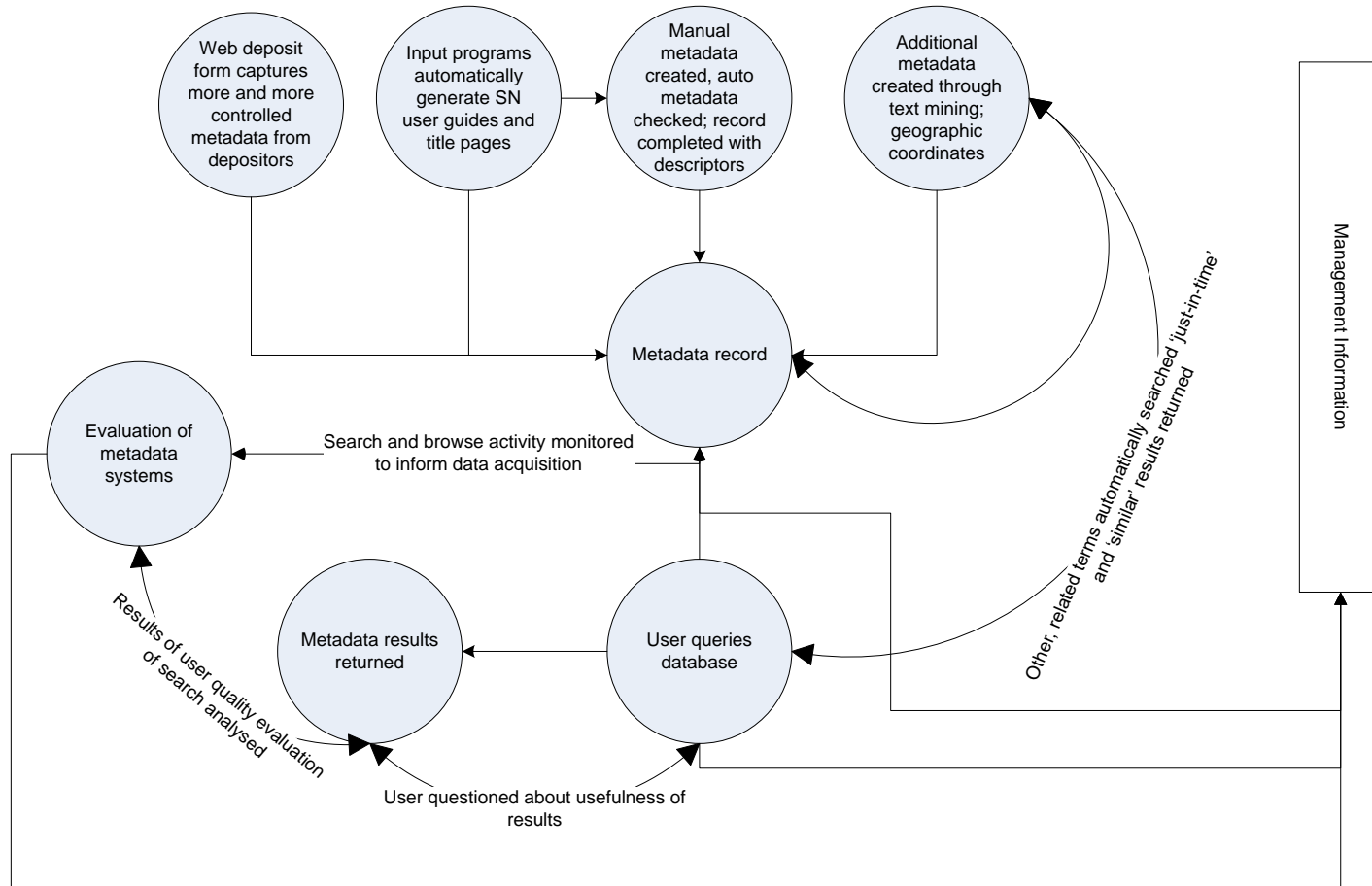
Conclusion

- JISC Intrallect report; quotation from Vic Lyte:

  - "A new researcher wishing to approach scholarly inquiry to determine the impact of global warming on penguin populations in South Antarctica doesn't walk up to a Librarian and shout 'Penguins!'."

  (Duncan, C. & Douglas, P., (2009). *Automatic metadata generation: use cases and tools/priorities*. Intrallect (for JISC): 2009)

References

- Aitchison, J. (1977). *UNESCO Thesaurus*. Paris: UNESCO.
- Broughton, V. (2006). *Essential thesaurus construction.* London: Facet, pp. 116-117
- Brown, S., Ross, R., Gerrard, D., Greengrass, M. & Bryson, J. (2006). *RePAH: a user requirements analysis for portals in the arts and humanities*. Leicester: De Montfort University.
- CESSDA. (2011). CESSDA : Council of European Social Science Data Archives. Retrieved from http://www.cessda.org/
- Connaway, L.S. & Dickey, T.J. (2010). The digital information seeker: report of findings from selected OCLC, RIN and JISC user behaviour projects. London: HEFCE.
- Duncan, C. & Douglas, P., (2009). Automatic metadata generation: use cases and tools/priorities. Intrallect Ltd (for JISC): 2009.
- UK Data Archive, University of Essex. (2011). Find data: our HASSET thesaurus. Retrieved from http://www.data-archive.ac.uk/find/hasset-thesaurus
- UK Data Archive, University of Essex. (2002). LIMBER: Language Independent Metadata Browsing of European Resources. Retrieved from http://www.data-archive.ac.uk/about/projects?id=1653
- UK Data Archive, University of Essex. (2006). MADIERA: Multilingual Access to Data Infrastructures of the European Research Area. Retrieved from http://www.data-archive.ac.uk/about/projects/past?id=1633
- Proctor R, Williams R & Stewart J. (2010). If you build it, will they come? How researchers perceive and use Web 2.0: a Research Information Network report. London: Research Information Network.
- British Standards Institute. (2005). Structured Vocabularies for Information Retrieval: Guide: Part 1: Definitions, symbols and abbreviations (BS 8723-1:2005); Part 2:Thesauri (BS 8723-2: 2005). London: British Standards Institute.

CONTACT

UK DATA ARCHIVE
UNIVERSITY OF ESSEX
WIVENHOE PARK
COLCHESTER
ESSEX CO4 3SQ

T +44 (0)1206 872001
E info@data-archive.ac.uk
www.data-archive.ac.uk

UK **DATA ARCHIVE**