Seemingly gliding: the power of metadata in academic resource discovery systems

Lucy Bell and Anat Vernitski

UK Data Archive, University of Essex

Abstract

Creators of resource discovery systems are at a turning point in development: their systems must satisfy the users' needs for the delivery of high-quality, academically-rigorous content, while also providing the most straight-forward, Google-like interfaces possible. The role of metadata in this scenario becomes paramount as users enter a world which *seems* serendipitous, but which is still subject to academic rigour, relevance and comprehensiveness. Ironically, such a system, necessarily supported by high-quality metadata, increasingly conceals them, thus lowering their profile in the user's mind, and their importance becomes even more camouflaged to the user community. Without them, however, there is no system.

This paper explores this dichotomy, describing a vision for resource discovery which creates a seamless journey between related, yet disparate, resources. It moves on to look at a case study of a single tool: the HASSET thesaurus, developed at the UK Data Archive. This tool's journey from an inhouse-developed thesaurus to the lynch-pin of many online services is described in detail. Lastly, the paper takes the theory of achieving focused results from a simple interface and, using HASSET, applies it to an existing service: the Economic and Social Data Service (ESDS). Metadata is shown to be the essential – and yet increasingly invisible – force at work in effective resource discovery systems.

Introduction

Looking at swans on a lake, they seem to be gliding effortlessly; however, this graceful movement is enabled by focused and powerful paddling under the surface of the water. Similarly, much metadata employed in resource discovery systems today is hidden, acting behind the scenes of the search interfaces that they support. Their role is absolutely vital, however, to the effective functioning of resource discovery tools. In many ways, the role of metadata as the power behind the search engine is of even greater importance now than ever before. It is ironic then, that these powerful, high-functioning metadata should be increasingly invisible.

Trends in information-seeking behaviour

The work involved in effecting efficient information retrieval has shifted over the last ten years, with both its location and its main players having changed, even if the associated tools, in many cases, remain the same. Ten years ago, to interrogate an online database effectively, users would have been encouraged to consult a thesaurus in advance, collating all the relevant keywords and constructing a complex search string, perhaps in consultation with an information professional. Today, that parcel of work, the essential task of unpicking what is required from the search, has moved behind the scenes, with those same thesaurus terms being accessed to a much greater extent than ever before by automated means.

Part of the reason for this change is, of course, the increased functionality of the technology available. Even if they do not count themselves part of the Google generation, many users are familiar with approaching a single search box and entering any terms of their own choosing. Convenience has risen in importance in the information-seeking process:

> 'Ease of discovery and access in getting to the information resources relevant to their needs, and in keeping themselves informed of events and publications in their fields, is critically important for researchers.' (Proctor, Williams. and Stewart, 2010, p. 35)

This quotation from the RIN report on researchers' approaches to Web 2.0 highlights the need for speed and simplicity in gaining access to data and other information; however, the report also indicates the scepticism with which tools such as blogs and wikis are viewed within the academic community. The expectation is still that academically-professional resource discovery tools will contain quality-assured, sustained and, in the case of journals, peer-reviewed content.

Internet search engines are simultaneously liked and mistrusted. The 2006 RePAH report highlights the fact that academics find the speed and simplicity of internet search engines to be appealing, while at the same time recognising their limitations (Brown, Ross, Gerrard, Greengrass & Bryson, 2006). A worry lurks that something key will be missed.

A dichotomy exists then: many academic users would like, indeed expect, fast and easy search interfaces; however, they also insist on their data discovery systems being of a very high standard. This emphasises the need for extremely high-performing metadata. The recent, JISC-commissioned meta-analysis of user behaviour projects supports this:

> 'Regardless of age or experience, academic discipline, or the context of the information need, speed and convenience are important to users … They also desire enhanced content to assist them in evaluating resources.' (Connaway & Dickey, 2010, p. 4)

Users are *expecting* the technology underpinning the systems to assist them in their resource discovery.

The JISC report is clear that 'library systems need to look and function more like search engines, i.e., Google and Yahoo, and web services, i.e., Amazon.com, since these are familiar to users who are comfortable and confident in using them.' and that, partly because of this, 'high-quality metadata is becoming more important for discovery of appropriate resources.' (Connaway. & Dickey, 2010, p. 5)

A vision of controlled serendipity

Part of the attractiveness of Google is, of course, the relevance of its results sets. Creators of resource discovery systems are currently facing a challenge. It's almost as if users wish to enter a virtual library and have the relevant books jump off the shelves and run to them.

The role of metadata in this scenario becomes paramount as users enter a world which *seems* as serendipitous as possible, while is still subject to academic rigour, relevance and comprehensiveness. Information providers need, perhaps, to establish a system of controlled serendipity. Such a system would be acutely dependent on high-quality metadata and yet, ironically, would also increasingly conceal those metadata, thus raising the potential for their importance to become camouflaged;

without the metadata, however, there would be no system at all. In order to provide effective and intuitive services, information providers must rely on well-developed and trusted metadata tools, some of which may have been many years in the development.

HASSET as a resource discovery tool within the UK Data Archive

One of the most powerful metadata tools behind academic information retrieval used at the UK Data Archive (the Archive) is the Humanities and Social Sciences Electronic Thesaurus (HASSET). This is an inhouse-developed thesaurus which enhances the catalogue-searching experience of users coming to look for data provided by the Archive.

HASSET has been developed by the Archive over the past 30 years or so. Initially, HASSET was based on the UNESCO Thesaurus compiled by Jean Aitchison in the 1970s (Aitchison, 1977). In 1979, the Archive approached UNESCO to receive a machine-readable copy of the thesaurus, which was duly received in 1980 in the form of two magnetic tapes. In the early stages the thesaurus used by the Archive was identified as a local extension of the UNESCO Thesaurus, with added terms and hierarchies. Originally, terms were added manually on card indexes, but later integrated as a computerised system.

Over time HASSET has been continuously expanded and updated for use as an online retrieval system; it still acknowledges UNESCO Thesaurus as the original source but it has now become an independent product. A multidisciplinary thesaurus developed primarily to support the UK Data Archive collection, HASSET reflects in its coverage the subject content of the Archive’s holdings. Coverage is 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 coverage continues to be developed as the holdings grow.

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. The latter use branches out beyond the remit of the Archive, as HASSET is used by a growing number of archives, libraries, research projects and government bodies in the UK and abroad.

Indexing using HASSET

HASSET is used by cataloguers and processors of data deposited and preserved in the Archive to index datasets compiled into studies. Keyword indexing is done in conjunction with cataloguing. The Archive’s collection includes both quantitative and qualitative studies and both are indexed with HASSET keywords.

As a general rule, for quantitative studies a keyword is assigned to each variable, or to each question or group of questions in a survey. Qualitative studies tend to be indexed at the level of 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.

Cataloguers are given initial training in cataloguing and indexing, as well as refresher training as required, within the Archive. Regular meetings for cataloguers are held monthly in order to discuss good practice, updates and queries, including those related to indexing. In this way the Archive ensures quality use and implementation of cataloguing and HASSET indexing.

The development of HASSET

New HASSET terms are developed regularly as required in order to cover the ever-growing collection of the UK Data Archive. Most new terms are initiated by queries of processors/cataloguers who flag up the need for new vocabulary to index a particular study. These queries are logged on a dedicated online internal Help Desk which allows for tracking and searching. Resource Discovery staff review each request in the context of the collection as a whole. It is important to encourage cataloguers' input regarding the need for new terms, as they are the front-line staff, who have the best insight regarding the data. Naturally they would be focused on the particular datasets they are processing, and this is where a more inclusive view of the whole collection is required, making sure that terms have a suitable scope for the collection as a whole and the correct level of specificity.

The creation of new terms in HASSET also involves much attention to the correct use of hierarchies. As Vanda Broughton explains when discussing the conceptual structure of thesauri, 'the major relationships are those of hierarchy' (Broughton, 2006, pp.116-117). HASSET uses all the principle thesaural relationships listed by Broughton, including BT (broader term), NT (narrower term), RT (related term), USE (from non-preferred term to preferred term) and UF (from preferred term to non-preferred term), and it additionally uses the relationship TT (top term above BTs). Considering the place of a new term in the hierarchy is one of the main considerations taken into account when creating new terms.

Finally, existing terms and hierarchies are also examined regularly and, when resources allow, legacy work is conducted in order to improve them and bring them up-to-date. This is done in order to serve better the growing collection of scientific and scholarly research upon which the Archive's collection is based.

In the creation of terms and hierarchies appropriate standards are used, namely BS 8723-1 and BS 8723-2 (Structured vocabularies for information retrieval), as well as relevant guides, specialist dictionaries, encyclopaedias and geographical gazetteers, and available scientific and scholarly literature. Terms created are documented, and Archive staff are regularly notified of any changes to HASSET.

HASSET as a source for multilingual thesauri

HASSET has already been extended once; it is the source for the multilingual European Language Social Science Thesaurus (ELSST), which was developed as an enhancement to HASSET with partial funding from two EU-funded projects - Language Independent Metadata Browsing of European Resources (LIMBER) (UK Data Archive, University of Essex, 2002) and Multilingual Access to Data Infrastructures of the European Research Area (MADIERA) (UK Data Archive, University of Essex, 2006).

The LIMBER project, which was conducted during 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. The data were coded, alphanumeric data from many thousands of social studies; the data themselves did not require translation, but the metadata, including the fields or code values used to interpret the data, did. A multilingual thesaurus of social science terms was constructed. A metadata model was developed using the World Wide Web Consortium (W3C) standard for metadata and the Resource Description Framework (RDF), to allow the construction of semantic definitions of terms in the thesaurus, making them more interpretable by users, and providing a standard basis for query and retrieval tools. Tools were also developed to construct and maintain the metadata.

Following on from LIMBER, the MADIERA project took the development of multilingual thesauri based on HASSET even further. Between 2002 and 2006 the project aimed to create an effective operational web-based infrastructure for the European social science community, which contains a wide range of data and resources from a number of providers, including members of the CESSDA[1] community. The resulting tool currently includes English (source), Danish, Finnish, French, German, Greek, Norwegian, Spanish and Swedish terms.

HASSET in use beyond the UK Data Archive

Already in the early 1980s, the UK Data Archive was approached by outside organisations who were interested in using HASSET for their own purposes. At that stage permission needed to be sought from UNESCO. In 1997 the Archive received ESRC funding which enabled it to develop the online version of the thesaurus further. This led, with UNESCO's permission, to HASSET becoming acknowledged as a separate product. Since then, HASSET continues to be used by organizations outside the 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.

The UK Data Archive is now interested in developing further its relationship with external users of HASSET, aiming for better communication regarding HASSET updates and development and for the sharing of good practice. Recently, the first phase of a user consultation project has been implemented in the Archive, and it is hoped that further developments in this area would enable both the Archive and organizations that use HASSET to benefit more from their relationship and their use of the thesaurus.

Resource Discovery in the UK Data Archive and ESDS

The UK Data Archive's main role is as the curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. With several thousand historical and contemporary datasets relating to society, the Archive is a vital resource for researchers, teachers and learners. It is an internationally-acknowledged centre of expertise in the areas of acquiring, curating and providing access to data. The Archive acquires high quality data from the academic, public, and commercial sectors, and provides continuous access to these data while also supporting existing and emerging communities of data users.

Access to these archived resources is made available via a number of other, Archive-managed services. The biggest one of these is the Economic and Social Data Service (ESDS)[2], the UK's flagship portal for research resources, which makes available key national and international survey data and qualitative data.

[1] CESSDA. (2011). CESSDA : Council of European Social Science Data Archives. Retrieved from http://www.cessda.org/

[2] UK Data Archive, University of Essex. (2011). Economic and Social Data Service. Retrieved from http://www.esds.ac.uk/

The Archive has under its control many resource discovery tools. From May – October 2010 a review was conducted of these tools in order both to map them and to make recommendations for improvement. This work was undertaken in the contexts of a move to DDI 3.0 and of bringing the Archive's tools closer to the vision of assistive interfaces, to help the users to focus their searches. Both existing and powerful metadata tools, such as HASSET, as well as new techniques and technologies, were explored.

The review described the 21 interfaces that the Archive manages. To its advantage, the development of so many tools shows the Archive's commitment to improving the user experience by offering many different routes into the data. Although these resource discovery tools are many and varied, a few already contain links between each other (such as the RELU-DSS Knowledge Portal which not only references its own metadata but also provides a path into the Data Catalogue). The review recommended that these metadata mappings should be further developed, and the resource discovery tools streamlined. This would both reduce the number of interfaces available to the user and also provide the opportunity for seamless journeys through the data, taking the user from catalogue record to dataset to variables to related publications and outputs and beyond.

The vision is one of a series of paths through the system so that, no matter which entry point is chosen, the user always has the option to visit one of the others and is always able to leave with what they want. This would be a large undertaking, linking studies in ESDS to publications and citations, and linking metadata from within the Data Catalogue to metadata in existence in other systems, including the website. It is also expected that faceted browsing will be implemented at the results stage.

The crucial element to the success of this work will be the connections made between the underlying metadata, in particular using HASSET to index further collections. Further automatic metadata generation will also be investigated, via concept parsing and relationship analysis; however, existing tools will continue to be used and may also be augmented to ensure that richer metadata are applied. It is hoped that using the same thesaurus across services and interfaces will go a long way to ensuring the success of this work.

Conclusion

Most academic users have become familiar with the focused functionality of Google searching. They want the freedom to enter terms of their choosing into search interfaces and yet still to return accurate and relevant results. As an information provider, the UK Data Archive is investigating how to point its users to the right results.

The JISC Intrallect report on automatic metadata generation contains a quotation from Vic Lyte, which sums this up nicely; Vic compares the use of a search engine search box with a human conversation:

> '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 & Douglas, 2009, p. 32)

Sadly, this is effectively what many searchers do when interrogating online databases. What the developers of digital information resources need to do today is to ensure that, on hearing that cry, their systems respond to the users with intelligent, metadata-driven suggestions to satisfy a more comprehensive search.

The Archive has, at its fingertips, one of the most useful metadata tools any information provider could wish for: a service-specific thesaurus. This tool is used to index data at an extremely low-level of granularity. Its further development, combined with other forms of precise indexing, will continue to augment the resource discovery tools managed by the UK Data Archive for many years to come.

References

Aitchison, J. (1977). *UNESCO Thesaurus*. Paris: UNESCO.

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.

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.

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.

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