# The roles of thesauri, metadata and automated keyword generation in document repositories

Marcus. R. Wigan[1]
Professorial Fellow, Australasian Centre of Governance and Management of Urban Transport,
Faculty of Architecture, The University of Melbourne, Victoria Australia
Phone +61 3 94599671
Fax +61 3 9459 8663
Email oxsys@optusnet.com.au

Robert Kukla
Research Fellow
Transport Research Institute and School of Computing
Napier University Edinburgh
10 Colinton Rd
Edinburgh Scotland EH10

Tel +44 131 455 2
Fax +44 131 455 2
Email r.kukla@napier.ac.uk

**Abstract**

Web access has transformed the way in which information can be discovered and accessed, but the questions of quality and reliability remain. Changes in the mode of publication and the Open Access movement go only part of the way towards addressing these now more urgent issues. This paper explores the interactions between repositories, document and data, discovery and accessibility from the point of view of researchers themselves, based on a researcher built repository framework and the practical solutions to metadata entry, access and the key area for metadata generation and input. Most of the current literature focuses on a library science approaches and communities, and the present paper comes to a more convergent view -but from the stance of transportation researchers. More effort is needed to ensure that the full benefits of these new resources are secured by greater equality and collaboration in interworking between library and transportation research practitioners. The positive initial results from automated keyword generation by using multiple transport thesauri within full text repositories now offers a real opportunity to improve these thesauri in a realistic interactive team environment- and in a reasonable time. Better transport analysis and modeling methodology keyword metadata could substantially improve document discovery in this area. A key issue given that metadata based discovery is fundamental for OAI-PMH and Z39.50 based repository information interchange and access processes.

[1] Emeritus Professor Napier University

## INTRODUCTION

The growth in research document repositories around the world has been driven mainly by library specialists, rather than end users or creators of the materials that comprise such repositories. Similarly, the Open Access movement has been driven by researchers with concerns about rapid and ready access to up to date materials, and worries about the constraints and delays inherent in many professional journals. Publishers, pressed by this two pronged squeeze on their previously eminent domain have moved both to even wider and more stringent transfers of rights from authors in return for publication, and at the same to various levels of accommodation to self archiving of materials in publicly accessible document repositories by authors.

The present paper is written by a researcher who needed a document, data and geospatial repository for real time use within a single project *(1)* and to deliver the outcomes as a repository at the end.

The project was targeted at the barriers to railway integration in the European Union, and looked at a wide range of aspects of the organizational, technical, spatial, social and spatial aspects of railways systems in the European Union, and thus required many different types on information be brought together and made accessible, and be supported with good communications.

The system developed for this purpose (the Napier Knowledge Base System or N_KB_S *(2))* proved to be very effective in managing the diverse materials required for such an ambitious repository, but the major problem encountered was securing the input of metadata into the system with the documents and data incorporated.

The present paper is a brief summary of experience with experiences during, and approaches developed after, the REORIENT project to find ways to address this situation. Information discovery is a key issue in all forms of repository and is manpower intensive.

## METADATA IN A MIXED REPOSITORY SYSTEM: INITIAL EFFORTS

To discuss automated means of addressing metadata generation before exploring the processes that led to this requirement is to miss key features of the process and how it is viewed by end users: the key players in this game.

The REORIENT project (www.reorient.org.uk) was a large project 6.3m Euros) aimed at identifying barriers to railway integration and freight movement market shares by rail in the eastern states of the European Union. One of the innovative aspects of this project was the commitment to develop a set of knowledge and communication resources to support eh project development and execution, instead of simply a repository of the formal deliverable outcomes at the end of the project (www.reorient.no).

This concept of a ‘knowledge base’ is a combination of document repository, data repository, geospatial information repository, thematic mapping tool, data cube analysis system, WIKI, controlled mailing lists and video conferencing tools, all presented in a unified single web portal format. The longer list of capabilities fully implemented in the project is detailed in *(1)*. Needless to say there was considerable surprise when most of the promised system was demonstrated fully operational less than half way though the project time period, and well in time for the incremental build up of documents, data and mapping to occur as subprojects were set up and produced results.

Clearly such a system had to handle multiple generations of drafts as well as final materials, and a search engine of considerable power and flexibility to be usable for the

multiple purposes of internal project communication and information sharing, updating and delivery. Also the groups who could have access to different sets of materials would have to change in membership and levels of access as the subprojects went through their lifecycles, so a sophisticated access and security strategy also had to be implemented.

User input was essential, so provisions for their managing their own security, metadata input, adjustment and editing were required.

The approach adopted to support the resulting REORIENT Knowledge Base (R.KB) for this assembly of materials in and for the REORIENT project was to develop a system design that could host a range of families of such Knowledge Bases. This was of Open Source tools, proprietary back ends for datacube handling and generalized document management, mapservers etc. This combination of tools was presented through a simple web portal interface. This infrastructure was termed the Napier Knowledge Based System. (N_KB_S), and this has also been used at a simple level to show how this underlying system design can be used to offer the same range of capabilities for other projects and project groups.

The WORLDNET project used this opportunity to explore end client reactions to how such a system could be used both for the same purposes as Reorient and also for collaborative online geospatial data input and editing (www.worldnetproject.eu). This took a single man week, and we invite readers to explore this themselves. The N_KB_S was also used to develop and deliver - within 3 days - a complete document repository for the 2.7Gb of unpublished papers from the last nine years of the European Transport Conference as a demonstration for the Association for European Transport (www.aetransport.org) who manages this conferences series, and which now has a formal Memorandum of Understanding for mutual recognition and collaboration with TRB. It is this latter implementation that is used as the illustration and trials for the present paper,

One of the issues addressed in this program was to make the access ot the document repository as simple as possible. To this end two facilities were implemented in the underling N_KB_S software system.

1. RSS feed. This allowed all activity on the relevant system to be communicated to those who wished to subscribe to the feed, and;
2. Z30.50 Protocol support (allowing such programs as Endnote *(7)* to access the document metadata directly without having to use the web portal at all)

The Z30.50 protocol is generally known only to the library and information science community, but it is also the protocol underpinning the incredibly widely used Endnote bibliographic software, which is licensed university-wide by hundreds of universities and research organisations and libraries (such as the Library of Congress) all across the world. This allows documents to be found by searching the metadata across the Internet, and the details to be downloaded in a directly usable form for citation, bibliography building and other uses by end users with no library science background at all. The N_KB_S supports Z39.50, and EndNote plugins have been set up (<20k in size) to allow any Endnote user to access the metadata resources in knowledge bases held in the N_KB_S

The choice of metadata elements in the N_KB_S was an important design criterion for the entire system. The basis selected was the 16-element Dublin Core, currently the most widely used metadata system for documents. The Dublin Core (so named for the location where the systems librarians met to define it) is under regular revision[2], but has proved a staple and enduring survivor as the basis for many document and data discovery systems.

Requiring the full 16 elements of the Dublin Core (see Fig 2 for the subset finally used) was swiftly found to be inadequate, as it omitted items that were expected by the users, and was too obscure and demanding for end users to bother with. The result of interactions with the users was twofold.

1. The provision of what appeared to the users like a set of folders in which to direct their materials
2. A reduced Dublin Core set (see Fig. 2) with some additional items derived from the submission process itself, limiting their workload

The pseudo folder provision was the result of the unease and lack of confidence and familiarity of the end users (who did their own submissions) in a pure search model of placement and retrieval of materials. The process of submission then became one of navigating an apparent series of nested folders to place the document where the user felt it should be (see Fig 5). Restrictions to a full free text (Boolean supported) search and metadata indexing approach proved to be too alien a discovery mode for almost all the users, and so this hierarchical apparent 'folder' approach as a support rapidly proved to be an essential feature, and encouraged users to make increasing use of the system as a whole.

The tools[3] used in the N_KB_S were able to extract metadata from Acrobat and Word documents, but using these proved to be impractical. Different versions of Acrobat and Word and different operating systems produced different metadata, so only a small subset could be used consistently. Efforts to explain how to create consistent metadata using the properties in Adobe and Microsoft software also proved to be unusable and unacceptable. Neither company offers any realistic ways of standardizing such metadata input, and the automated metadata is often wildly inconsistent between one operating system or release and another.

Yet a quick demonstration of the power of using metadata fields in searches immediately impressed end users. TREND was often used as it was the name of an associated consortium and also a generic word. Searches using the Google approach (we included a Google button in the system to allow people to discover the shortfalls for themselves) found 199 items, using the metadata found the three that were actually about or by the TREND consortium. A very familiar story to anyone using powerful Boolean searches, CCL, or metadata supported systems – but not exactly the daily fare of most transportation engineers.

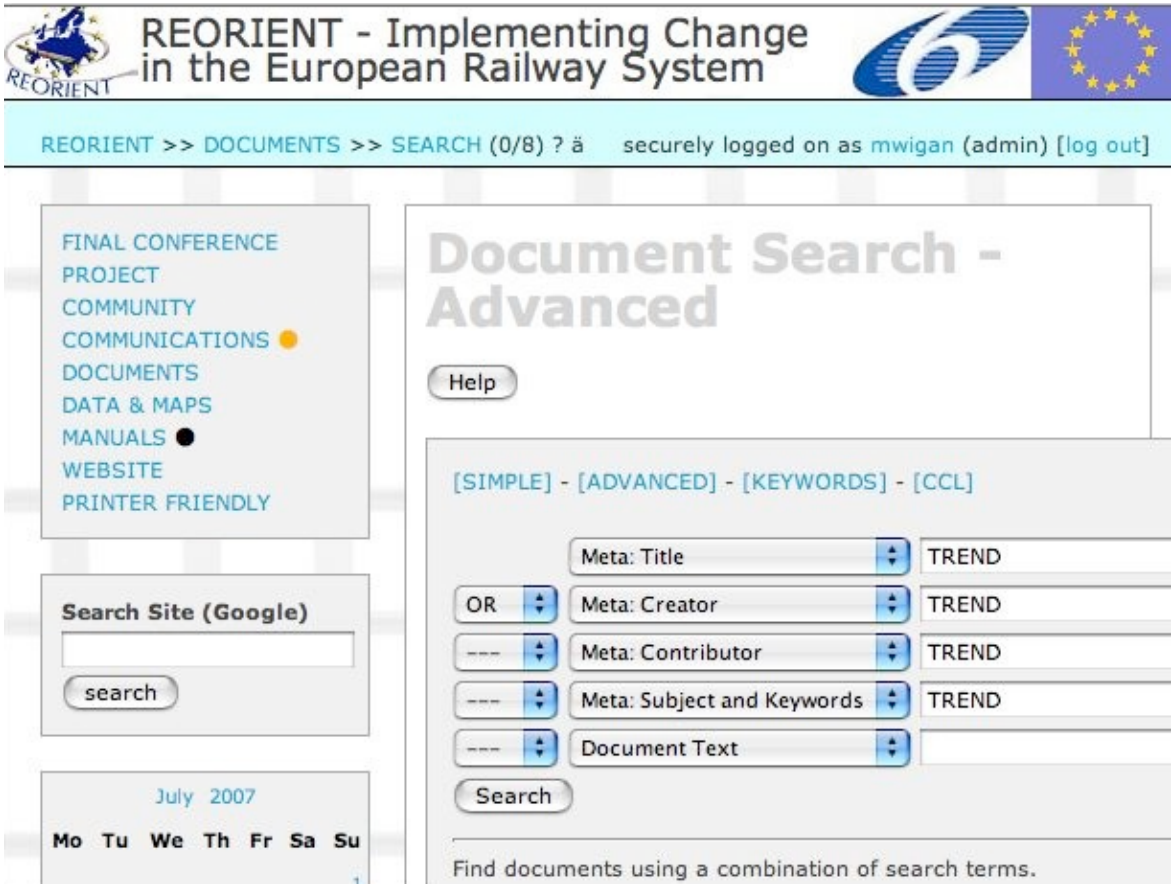
Fig. 1 shows the Reorient Knowledge Base using the Search functions customised by our team for the Reorient Knowledge Base implementation. The options for searching

[2] http://dublincore.org/

[3] SAIC;s TeraText and InOut systems, included inside the N_KB_S framework

are Simple, Advanced, Keyword and CCL (the latter being the international full search language: rarely used in most applications other than for the large scale intelligence and security areas where this very powerful document engine is widely used.

**FIGURE 1. Metadata only search in the REORIENT Knowledge Base online**

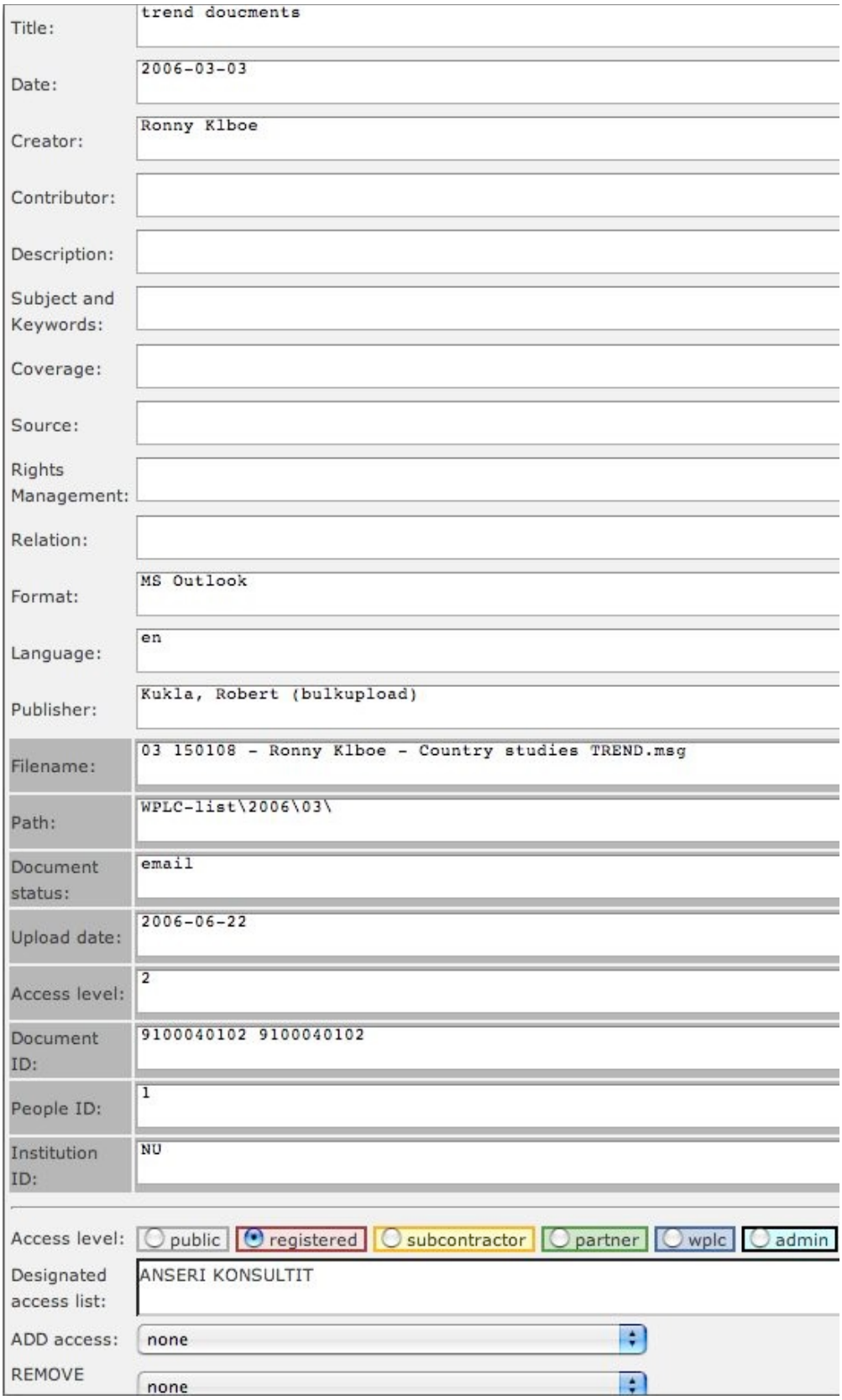



In Fig.1 the free text option at the bottom of the list is simply left blank and not invoked in the search. The query shown is limited to the appearance in the metadata of the target documents of with the word TREND in the Title of the document or listed as the Creator of the document[4]. This search produced the three documents being sought, with out the noise of the other 197 when using free text search on the entire document text body as one would in a Google or Simple search

The results of demonstrating the power of metadata-only searching had an unexpected, and highly undesirable effect. The end users immediately appreciated the power of such searches – and then expressed guilt at not putting in metadata in their own documents. This guilt response also extended to seminar invitations, when Georgia Tech subsequently expressed a wish for a seminar on a subject of my choosing- "as long as it isn't metadata".

**FIGURE 2. Mass Search outcome metadata editing and access control**

[4] Title and Creator being two of the 16 Dublin Core items that were retained in the system



| | |
|---|---|
| Title: | trend doucments |
| Date: | 2006-03-03 |
| Creator: | Ronny Klboe |
| Contributor: | |
| Description: | |
| Subject and Keywords: | |
| Coverage: | |
| Source: | |
| Rights Management: | |
| Relation: | |
| Format: | MS Outlook |
| Language: | en |
| Publisher: | Kukla, Robert (bulkupload) |
| Filename: | 03 150108 - Ronny Klboe - Country studies TREND.msg |
| Path: | WPLC-list\2006\03\ |
| Document status: | email |
| Upload date: | 2006-06-22 |
| Access level: | 2 |
| Document ID: | 9100040102 9100040102 |
| People ID: | 1 |
| Institution ID: | NU |

Access level: ○ public ◉ registered ○ subcontractor ○ partner ○ wplc ○ admin

Designated access list: ANSERI KONSULTIT

ADD access: none

REMOVE none

While this is probably mildly amusing to read, it is a serious issue in library science as any lack of metadata insertion by end users is perhaps the most critical resource constraint to wider accessibility and visibility of materials in electronic holdings and depositories.

This response is not simply one of not wanting to put in metadata, but the uncertainties of how to do it efficiently and consistently. and an implied plea for support tools to make it easier and not a fresh burden on researchers.

The first stage of providing such support was to find a way to allow document owners to edit the metadata for groups of documents in one go, and to adjust who could (at that stage) see them anyway. This allows both a restriction in visibility to informed and closer workers, and also an easier opportunity to add material to the metadata fields as and when the document(s) become closer to a final form. Fig. 2 shows the way in which this was done.

Fig 2 is the screen that appears if a single document (or a search) has been undertaken and editing of the metadata of wither the single document or all those found in the searching were to be edited. In Fig.2 this is one of the three results from searching for TREND using the query in Fig, 1. This is an email message, as all email is automatically archived in the document repository at a high level of security. Access to these records in protected, and even the metadata may be viewed even only with a password.

Here the truncated Dublin Core list of descriptive is shown explicitly, plus the shaded extra fields automatically generated by the system. The Path is the pseudo folder address mentioned earlier, and the file name is the original file name of the uploaded document. The document ID is the unique identifier in the data base and is created at upload time.

The inducement to end users to use this process is the security management offered in the lower part of the Figure. Here single documents or entire sets of search results can at one stroke be changed in terms of access to different levels of security or public access. In addition a designated access list is maintained for all documents which can be modified using the lower two bars (which are drop down menus containing all the organisations and individuals registered on the Knowledge Base .

Organisational and individually registered parties on the Knowledge base can be added or removed by selecting one or other of these two fields, which will provide a popup from which can be selected the party required to given or debarred access. Only one can be done in each such process, but it can be repeated as soften as required. the list. In the example Anseri Consult (a subcontractor) has been added to the access list independent of their assigned overall access level to the Knowledge Base.

We would like to have been able to report that this was taken up with enthusiasm, but we cannot. These facilities were however heavily used by the knowledge base experts in the project to make the processes of metadata updating manageable within the workload of a very small team. The screenshot shows the automatically extracted metadata before any subsequent editing.

**METADATA IN A MIXED REPOSITORY SYSTEM: AUTOMATED METHODS**

Metadata is critically important in document and data discovery. Consequently, as the resources were not available to do more than what was enabled by the tools summarised in the last section and which was still resource intensive, better ways had to be found.

The documentation of the Napier KB System (N_KB_S) could not be done solely using REORIENT Knowledge base examples as the already rife confusion between the knowledge base (assemblage of materials) and the framework within which it is built would be all to easy to feed by doing this.

An opportunity was taken to create a separate knowledge base as a trial publication demonstration for the Association for European Transport, who organise the European Transport Conferences. These are managed by PTRC and are the successors to the previous series of PTRC Conferences that ran for over 25 years. Early in 2007, AET was mutually recognized by TRB and mutual exchanges of endorsement were in effect for the 2007 ETC Conference.

Consequently explorations of better ways to make materials more mutually visible are the subject of an AET Council Subcommittee Chaired by the corresponding author of the present paper. Currently AET makes the ETC papers available only to members of the association, on the AET website, and methods of increasing the visibility and utilization of this substantial resource are needed.

There is a major international movement to make document repositories mutually visible to each other, using a protocol (OAI-PMH *(3)*) which allows the metadata to be easily exchanged. The documents themselves remain resident in the original repositories, and it is only the metadata that is exchanged. The key features *(4)* for an effective repository to participate in the rapidly globalizing area are

1. An OAI interface (see OAI-PMH)
2. A unique identifier
3. Compliance with Dublin Core for metadata

These are also the compliance requirements for the German Initiative for Networked information (DINI) Certificate of Compliance for a repository *(4)*.

These requirements are still not familiar to professions outside the library sciences community, although there has been a steady build up of interest in metadata with in TRB, with a subcommittee established in 1998 and a high level subcommittee (of the Data Section as a whole) after a major review *(5)*. The remarkable feature of this effort is that the major efforts of library science in Open Access initiative were entirely missed. Understandably so, as the focus was essentially on data – the prime direct concern of most TRB participants.

The Napier KB Framework handles all the DINI requirements, but also covers data, data cubes, thematic mapping and geospatial data. As a result it bridges the two worlds of the end users in transportation and the library science infrastructures emerging as an incipient global standard. The emergent Data Observatories movement to manage data and documents in one framework for web access are another indicator that this type of repository is now needed by yet another major group: local and regional governments *(6)*.

This development is encouraging as metadata play a major role in both domains, but making the links between them have proved difficult. The popular ENDNOTE *(7)* bibliographic and reference management system actually uses Z39.50 to allow very fast searches on global libraries and to integrate the metadata thus located directly into personal bibliographies. This concept is very familiar to many researchers and practitioners, while the underlying metadata and communication standards are not. The N_KB_S supports Z39.50 access to its metadata, and this is indeed by the end user community.

A trial repository was set up for the AET to meet the Z39.50, OAI-PMH standards and also to provide for the layered accessibility required to cater for the variety of Open Access regimes: an issue addressed in the next section. It actually proves easier to set up 2.7 Gb (i.e. 90% of the ETC conference holdings) than be too selective, even for a trial. This was due to the design of the N_KB_S where minimum effort for input had been emphasized.

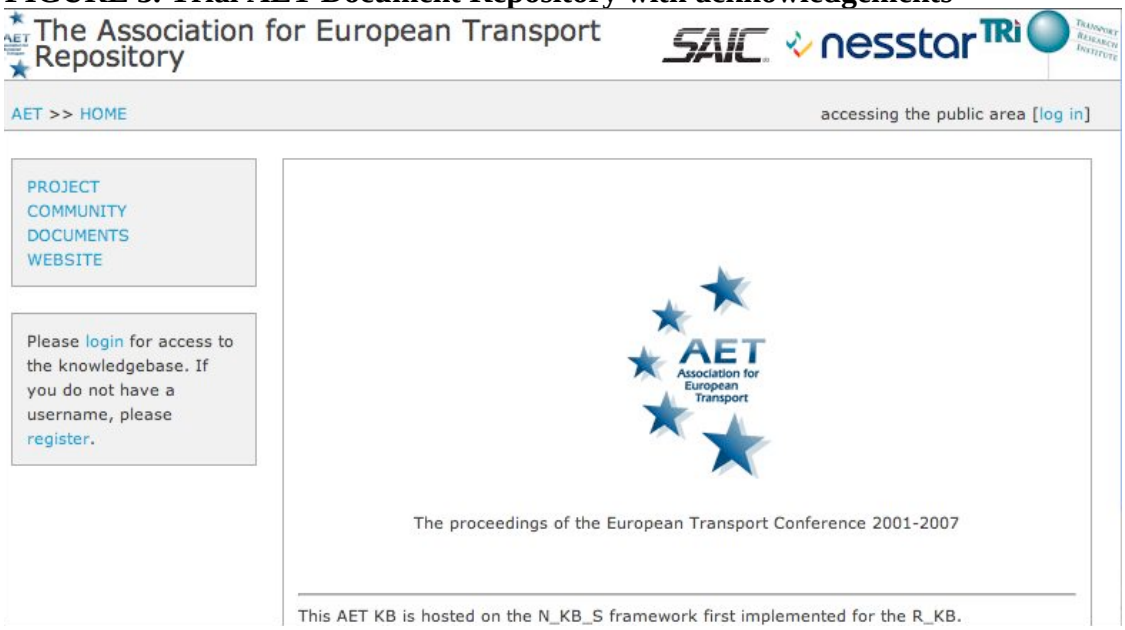
The entire test set up, access to documents, integration and opening to the web (passworded of course at this stage as it was a trial) took less that three days. The front page of this system is shown as Fig.3 (and is accessible at www.aet.reorient.org.uk)

Clearly the usual fine tuning and interface work would double or triple this, but it is an indication of the efficiency now available to build very large repositories of documents with both sophisticated and unsophisticated access facilities.

The data aspects take a bit more effort in the N_KB_S and this experience should not be taken to be typical of a wider range of materials to be held than simply textual documents.

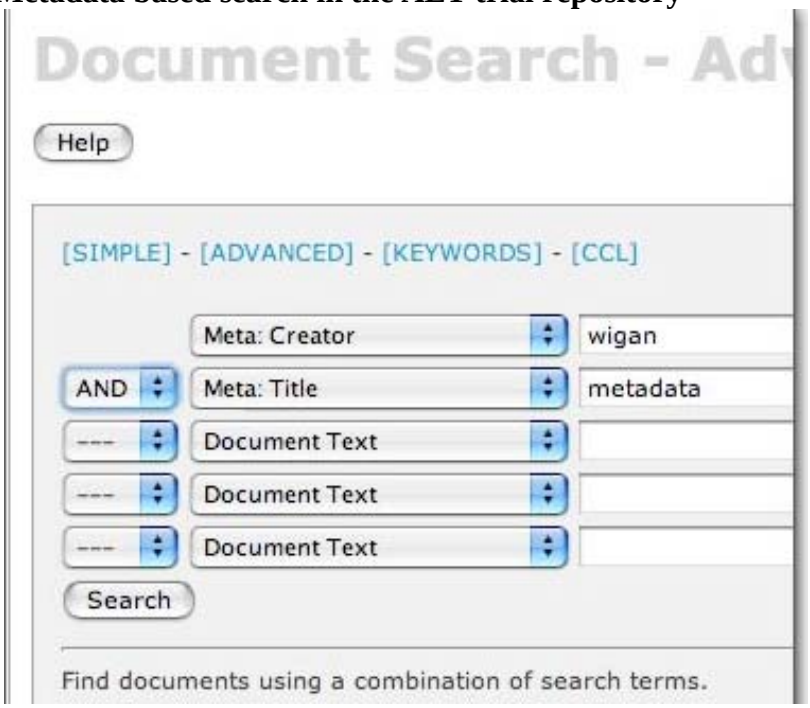
It must be emphasized that this trial repository was not at the time of writing any more than a trial, and was not endorsed as a public facility. This situation has subsequently changed, and access provision could be provided at the time of any presentation or publication.

**FIGURE 3. Trial AET Document Repository with acknowledgements**



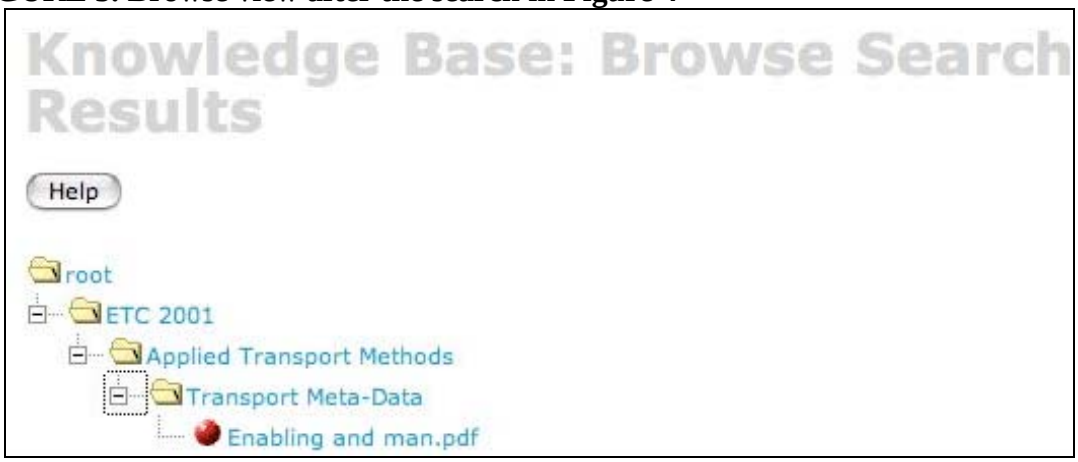
The next Figures (4,5) reiterate the search process summarised in Fig,1 for the reorient Knowledge Base, and adds the corresponding optional pesudo folder view (Fig, 5)**.** Fig. 4 specifies documents whose creator was “wigan” and whose title contained the word “metadata”.

**FIGURE 4. Metadata based search in the AET trial repository**




Switching to the pseudo folder mode (“browse”) shows where this document lies in the hierarchy.

**FIGURE 5. Browse view after the search in Figure 4**




With several thousand documents to deal with, manual metadata input was clearly impossible to create in a short space of time, certainly not justified for a trial, and probably not appropriate in terms of manual input resources for this scale of organisation.

To make this a manageable task, meeting the expectations of end users must be the best guide. In general users will not be expecting the finer elements of the Dublin core list of elements to be recorded – let alone searchable- so the clear option is keywords. This was the clue required to make some advances.

**ADRESSING AUTOMATED KEYWORDS VIA THESARURI MATCHING**

The prime source of considered keywords are thesauri. These are carefully selected dictionaries of words that are found to be useful in identifying document contents. There is a great deal of experience both in developing such thesauri and in making use of them. This experience is, however, now usually restricted to professional librarians or these cataloging documents or completing metadata descriptors as required when entering documents into library repositories.

The major users are librarians injecting metadata (as indeed keywords are the most commonly recognized form of metadata in most end user fields), and these dictionaries are carefully updated and regularly reviewed. In the case of the AET document collection the field is unambiguously transport, and so the two best Thesauri that we were advised to use were the Australian Transport Thesaurus *(8),* and the Transport Thesaurus developed by the US National Transportation Library *(9).*

Clearly both would be needed, as there are many Americanised spellings of English words, even in this specialised field. The use of both an Australian and an American Thesauri was expected to cover these variations at least in the first instance, and the recent revisions and updating of both would ensure that the terms coverage of both would be a good as one could reasonably hope for.

The document engine used in the N_KB_S is TeraText, which has its own Java-like programming language (ACE). This was used to set up a procedure to analyse all the words in he AET trial repository, and to match these to both thesauri.

The technique used was to check all the words in each full text record against both thesaurus dictionaries and collate the results as candidates for the keyword field of each of the documents examined.

First the entire document data base was matched against both thesauri, and all the words that occurred throughout every document, or which appeared only once. were pruned from the set of keyword candidates.

Each document was then searched and matched to this combined list of candidate keywords, and these words were then injected into the keyword field of the document concerned.

Needless to say there were a great many keywords for many of the documents. An example is given from the record (Fig 6) picked from the search of the AET repository shown in Fig.4. The set of keywords generated by this process are shown as Fig. 7.

Viewing the metadata (see Fig 8) will show both the metadata extracted automatically from the Acrobat files imported from the AET records, as well as the entries injected into the keyword field the result of the dual matching keyword generation process.

**FIGURE 6. Record view resulting from the search in Figure 4.**

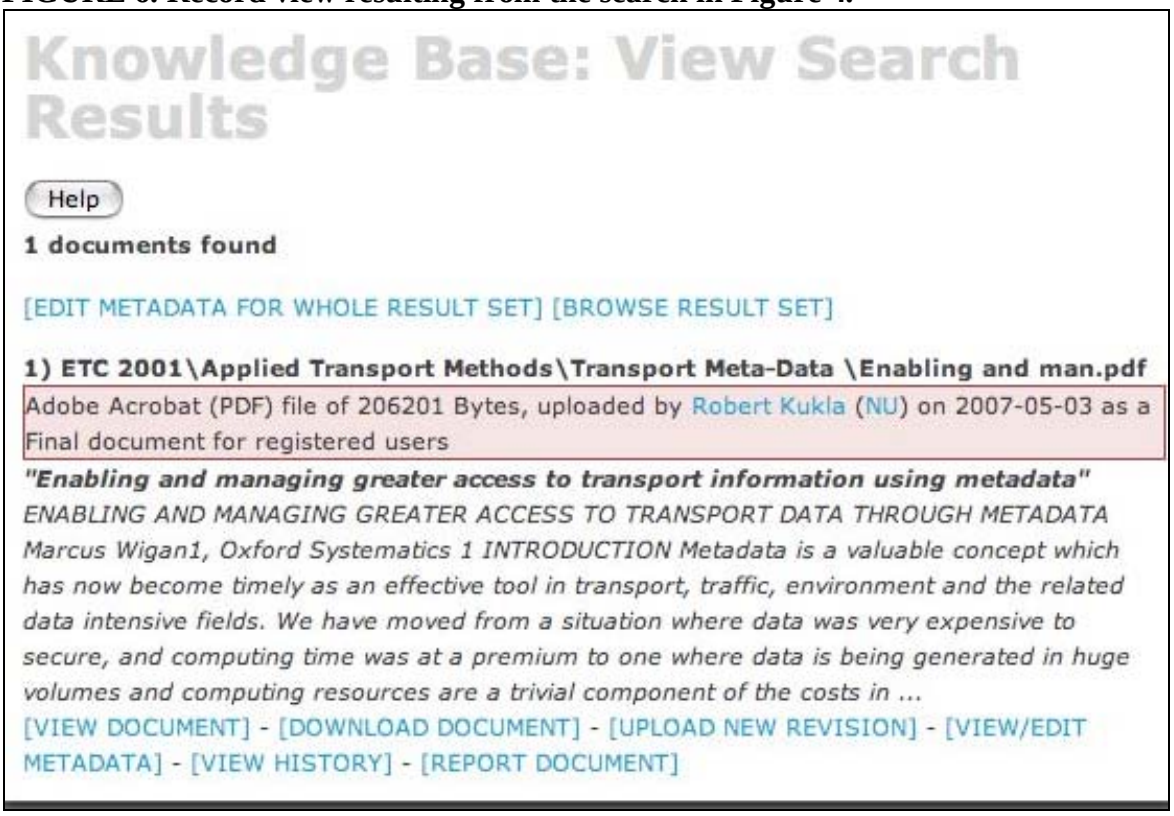



**FIGURE 7. Automated keyword generation results for the paper in Figure 6**

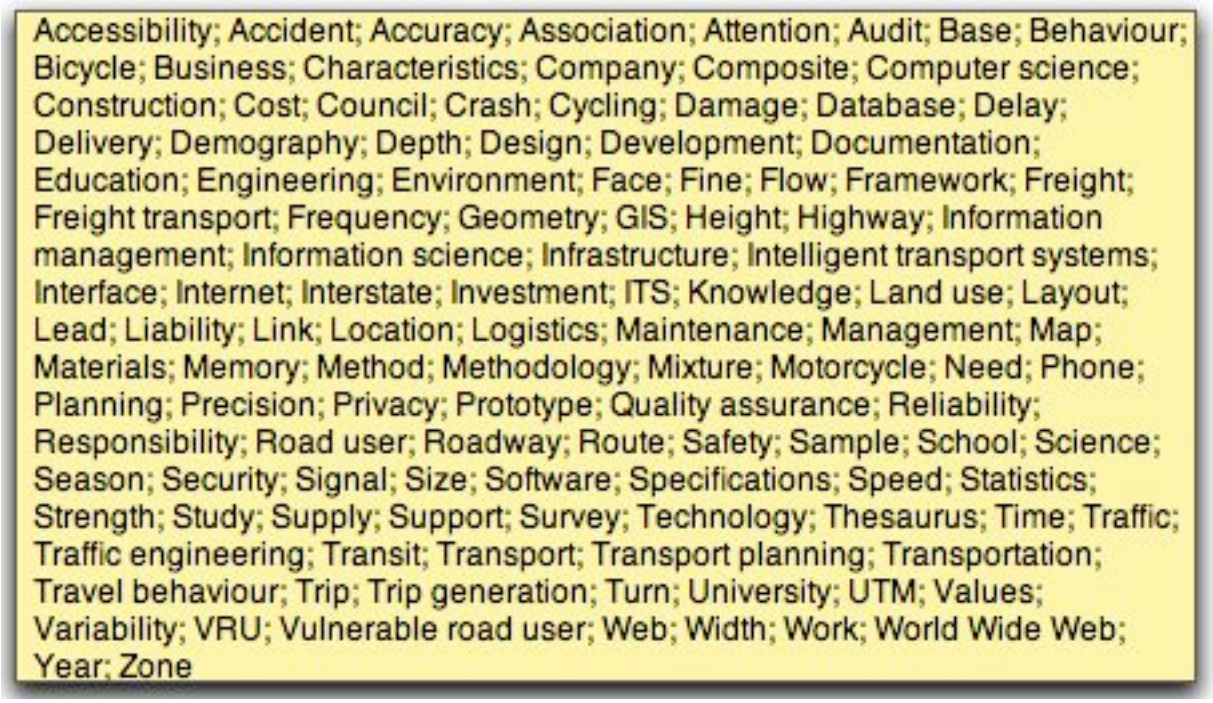
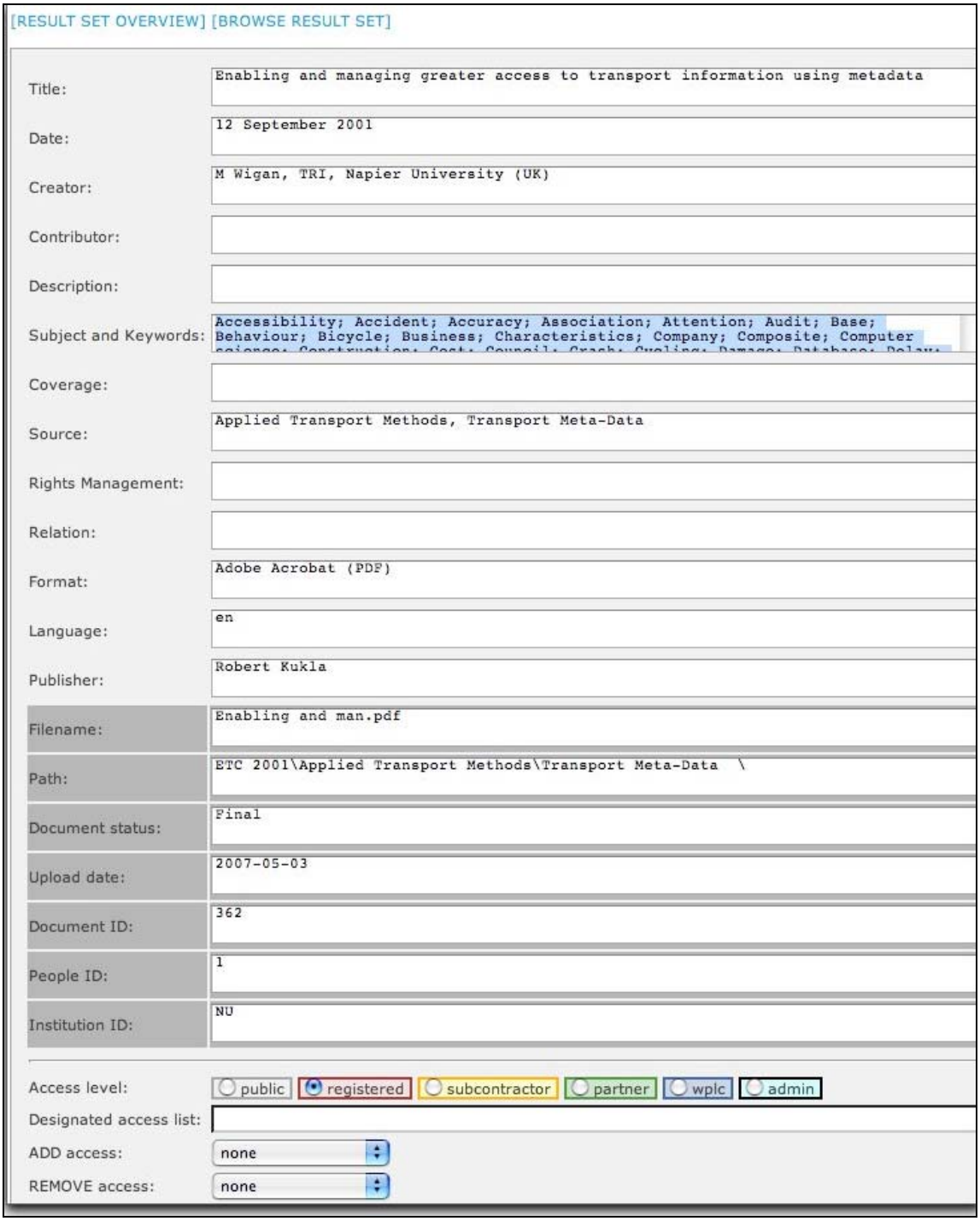



**FIGURE 8. Record view resulting from the search in Figure 4**

There are now many keywords. The results of the **Edit Metadata** command are shown in Fig.8, but the list is not shown in full as the scrolling window containing the

keywords can only be shown in a fixed position in Fig.8. The full set of keywords generated for this particular document are shown in Fig.7, and can be examined to consider the value of the keyword creation process.

The first question when viewing Fig 8 is: is this a useful outcome? Give that the other metadata available in the normal Acrobat file is so limited, the answer is probably yes.



However the value to the end user of such assemblages of keywords can only be tested by doing searches. A considerable number were done over a range of transport subjects and topics by a transport specialist, and there were two consistent conclusions:

1. The papers found were generally good matches to the expectations
2. That papers with a strong methodological content were very badly served, sometimes without a single automated keyword.

While the first finding is very encouraging, the second was half expected by the corresponding author. Methodological terms in transport thesauri have been poor ever since the first IRRD[5] keyword set was generated at TRRL in the late 1960's, and the corresponding author tried to add some even at that stage. This general issue remains a major shortfall in the utility to research users of keyword-based searches even after the intervening 40 years.

On consideration the outcomes are sensible. The thesauri from which the keywords were selected have been carefully built and refined to assist in searches, although with some strange encodings (one that applied for many years was the use of the word 'equation' in the IRRD thesauri to represent any kind of model being described or used in the document concerned). The matching across the full text ensured that any mention of appropriate words would be picked up anywhere in the document, and the steadily improving selection in these two thesauri would tend to make the selection more effective as the thesauri were updated.

This appears to be the case. Searches using these extended sets of keywords only appear to give good results when compared to the same searches done on the full text of the document – with the exception of methodological terms.

## CONCLUSIONS

Metadata perspectives from library science and end user data standpoints are not yet convergent, but the use of the carefully developed transport thesauri can assist when new transport document collections are put into repositories. This massively lowers the barriers in terms of time and effort for consolidating such materials, and is most encouraging.

Unfortunately the long known weaknesses of transport thesauri in terms of discriminating methodological coverage are apparently still in existence. We feel that using the automated tools that we already have could allow us to test the effectiveness of a range of such methodological terms by running our system over multi-gigabyte document collections (such as the AET trial) and experimentally determine what works.

This might seem to be a slow, unsophisticated and ponderous way of doing this, but we would contend that the automated processes are now so fast that in our initial experience this is probably the most efficient and fastest way of proceeding, as the interactive exchanges with specialist librarians will form an essential part of the process to develop and refine effective methodological keywords for transport thesaurus enhancement and use.

---

[5] The International Road Research Document database, coordinated by an OECD Road Research Program Committee and commercially available as the "Transport CD" from Silverplatter. For many years: now renamed the ITRD (International Transport research Document database)

We look forward to active collaboration with major thesauri in these trials, and hope that it might prove possible to establish a suitable range of collaborators through the TRB Library and Metadata related committees and subcommittees and link end user researchers and librarians to develop a suitable range of extended methodological keywords for general use as a result

## ACKNOWLEDGEMENTS

This work was undertaken at Napier University, led by the first author, subsequent to a project requiring a document data and geospatial repository supported by the European Commission Directorate of Transport and Energy (DGTREN) under the REORIENT project on barriers to railway integration.

## REFERENCES

1. Wigan, M.R, R. Kukla, A. Cannon, P. Grashoff, M. Benjamins A new data, document and geospatial repository: Knowledge base and project support for a major international railway project **[Submitted to OA08]**
2. Wigan, M.R, R. Kukla, A. Cannon, P. Grashoff, M. Benjamins. *The REORIENT Knowledge base: a knowledge base and project support for a major international railway project*. REORIENT Final Conference, Brussels, 31 May 2007. https://www.reorient.org.uk/pdfs/RKB_NKBS%20Confhandout.pdf (accessed 31 July 2007)
3. OAI Executive. *The Open Archives Initiative Protocol for Metadata Harvesting*. http://openarchives.org/OAI/openarchivesprotocol.html (accessed 31 July 2007)
4. Dobrantz, S., E. Mittler, H. Neuroth, P. Schirmbacher, and F. Scholtze. *Building an E-Publications Infrastructure*. Deutsche Initiative fur Netwerkinformation E.V. DINI Schriften 7-en, 2005.
5. Chiaio, K,. J.Hall, F.Harrison, J. Kreideweis, J. Koenemann, R. Gillmann, S. Tuener, M. Wigan and J. Zmud, *Transportation Metadata: Role of Data and Information Technology Section* http://www.trb.org/committees/datasection/DataSection-Metadata.pdf (accessed 31 July 2007)
6. Wigan, M.R. *Data Observatories and Metadata: linked issues in operation*. Data Stewardship session, TRB annual Meeting 2006.
7. www.endnote.com accessed 31 July 2007
8. Cox. L., Martin, A. and Meier, A. (2007) Australian Transport Index Thesaurus. ARRB group Vermont South. 130pp. http://www.arrb.com.au/documents/libraryThesaurus.pdf
9. http://ntlsearch.bts.gov/tris/trt.do accessed 31 July 2007