Improving access to transportation documents: the roles of repositories, thesauri and automated keyword generation

Marcus Wigan

Principal Oxford Systematics Australia

Emeritus Professor of Transport Systems, Napier University Edinburgh

Professorial Fellow, GAMUT Faculty of Architecture, The University of Melbourne

Summary

- Documents in transport rarely have good metadata
- Thesauri embody much effort and skill
- Transport documents can now be held in full text
- Inverted analysis of full text against thesauri can..
  - generate keywords automatically
  - be used to develop keywords to cover gaps

Context

- This paper and the repository system it uses was created by an active transport researcher
- The end user needs of researchers need attention
- Why is the gap between researchers and library science so large?
- How do we bridge it?
- Researchers inputting metadata is an additional overhead to them so… this paper

Transport thesauri and automated keywords

Where does resource metadata come from?

- Metadata value is rarely appreciated by end users
  - When shown its importance, the usual response is one of guilt for non-entry rather than enthusiasm
  - Users are now largely responsible for subject domain metadata input
  - Document repository operators find metadata input a major resource concern
  - Little use of specialist thesauri even by specialist librarians in such environments
  - No role for librarians in the progress of a project, as distinct from its final classification and holding

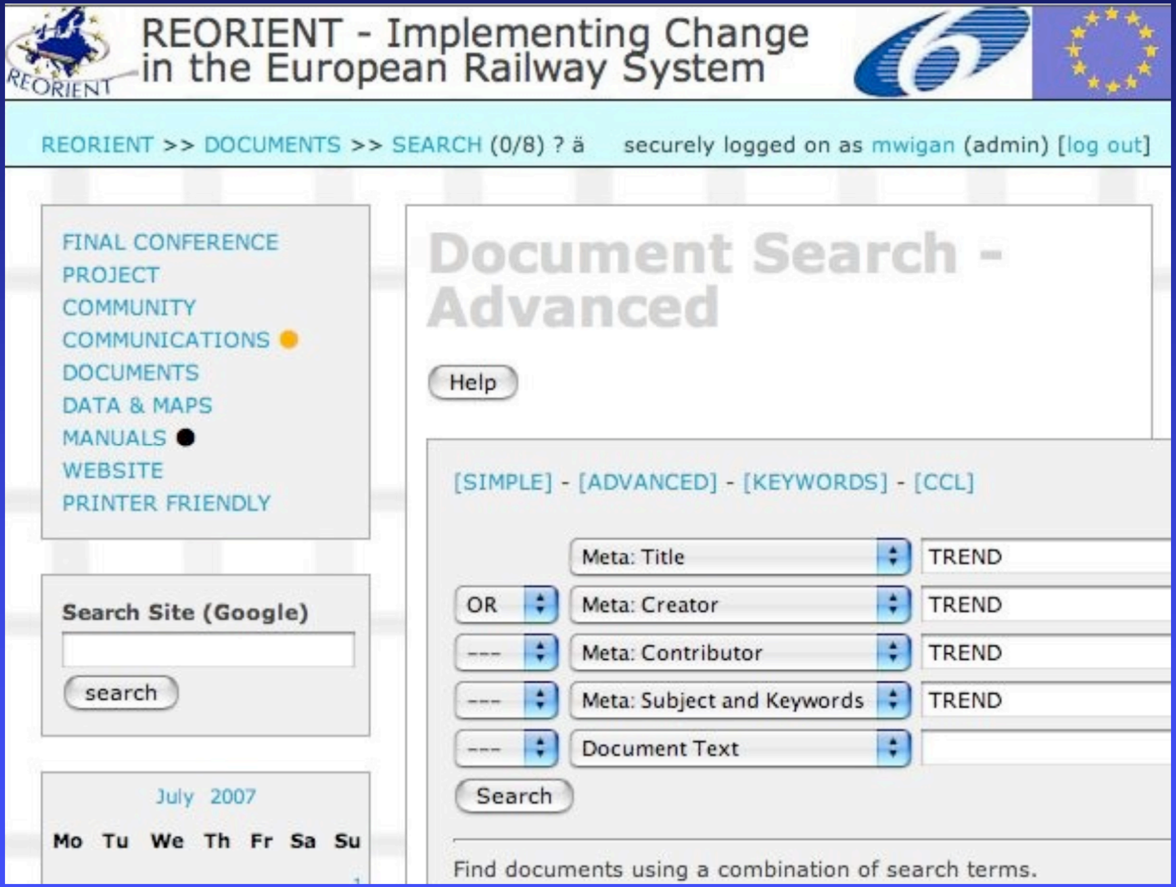
Context of this paper

- Written by a researcher who was desperate for a usable repository to support projects as well as deliver them.. but who could get nothing from the systems library communities
- The encompassing repository catered for documents in several generations, levels of security and progressively updated.
- Metadata input by end users very therefore very hard to secure
- The broader system in which this was important also handled data, geospatial information, dynamic mapping etc, and comprised a complete active Knowledge Base rather than just a simple document storage system

The Napier Knowledge Base System

- The software system developed for Knowledge Base building (initially for the ReOrient project)
- The resulting Reorient Knowledge Base very well received at the 2007 Freight Users Forum
- Includes a full SGML based document engine (SAIC's TeraText) - but needs middleware
- A most vexing issue was securing good metadata to allow efficient resource discovery in this very large resource
- Came to head with a Conference 'demo' (actually we built the whole 2.7Gb working repository in 3 days- as it took no more time than a limited demo)
- Lousy metadata….. And free text search is NOT enough

Searches need to include metadata



REORIENT - Implementing Change in the European Railway System

REORIENT >> DOCUMENTS >> SEARCH (0/8) ? ä securely logged on as mwigan (admin) [log out]

FINAL CONFERENCE
PROJECT
COMMUNITY
COMMUNICATIONS ●
DOCUMENTS
DATA & MAPS
MANUALS ●
WEBSITE
PRINTER FRIENDLY

Search Site (Google)

July 2007

Mo Tu We Th Fr Sa Su

Document Search - Advanced

[SIMPLE] - [ADVANCED] - [KEYWORDS] - [CCL]

	Meta: Title	TREND
OR	Meta: Creator	TREND
---	Meta: Contributor	TREND
---	Meta: Subject and Keywords	TREND
---	Document Text	

Find documents using a combination of search terms.

- Boolean searches including metadata fields

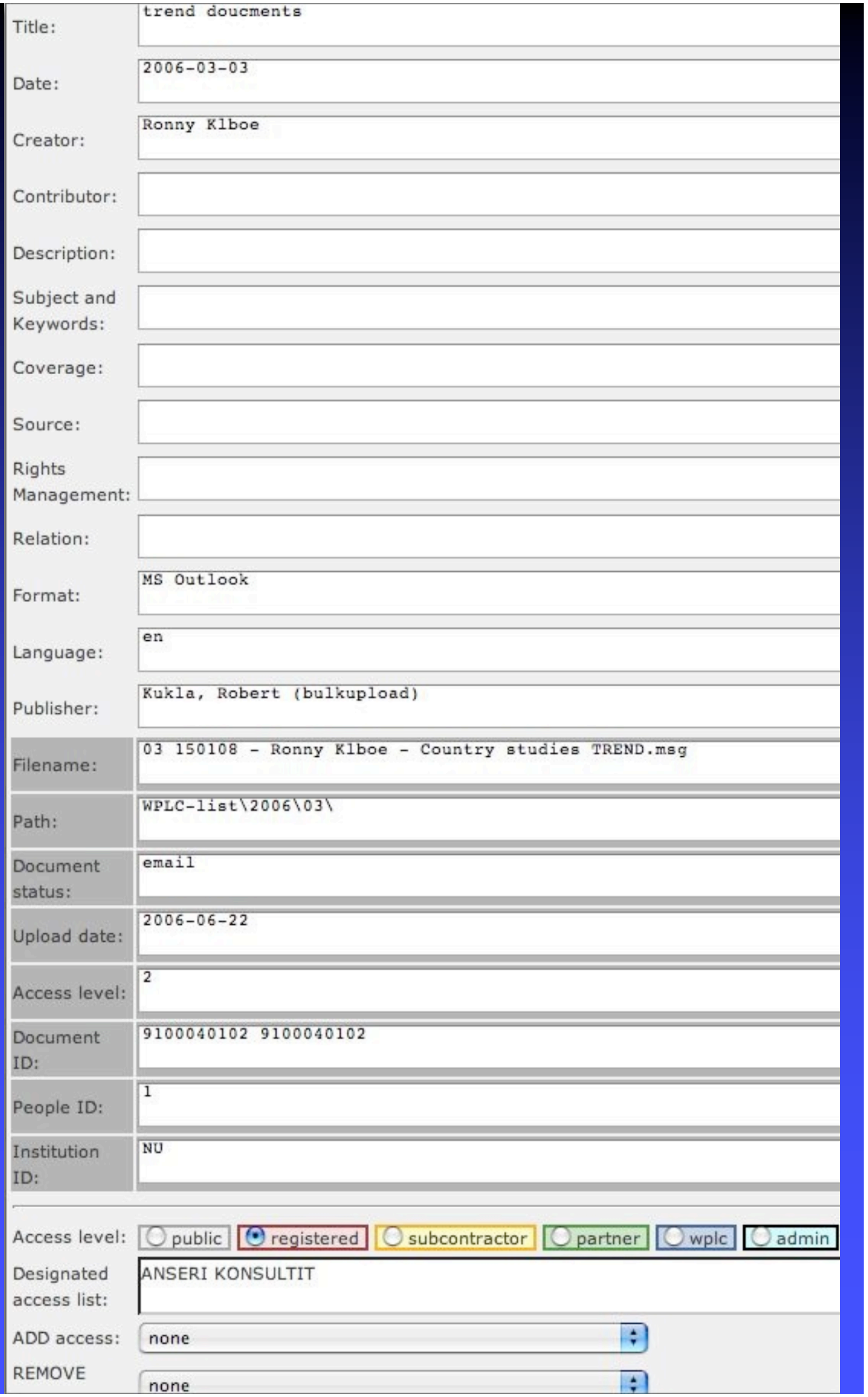
Transport thesauri and automated keywords

Which metadata fields?

Dublin Core Subset

Automatically Extracted Items

Access Control Fields



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

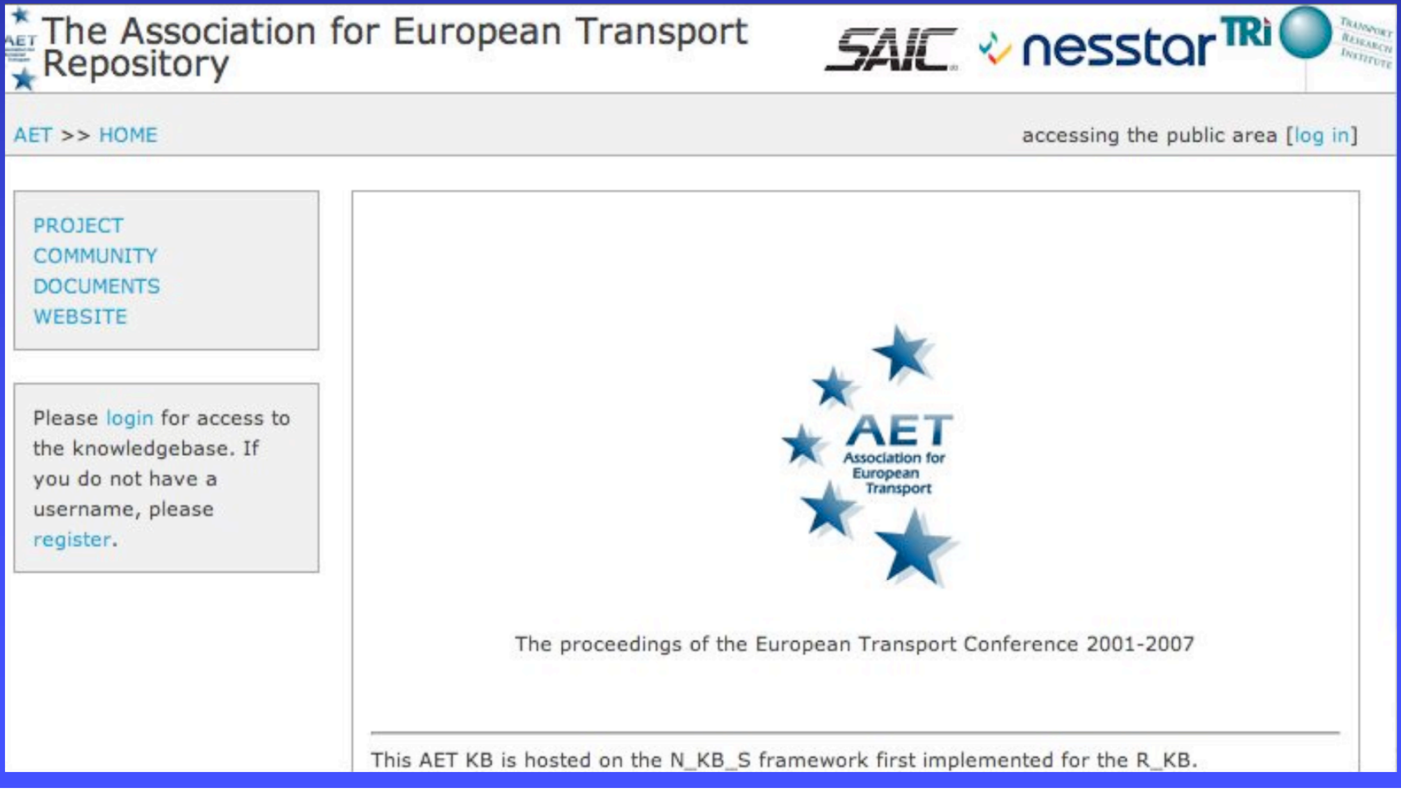
Designated access list: ANSERI KONSULTIT

ADD access: none

REMOVE none

Apply the Knowledge Base system to create a full document repository

- The European Transport Conference from 2001
- Front end and 2.7Gb of docs installed in 3 days



The Association for European Transport Repository

SAIC
nesstar
TRI
Transport Research Institute

AET >> HOME

accessing the public area [log in]

PROJECT
COMMUNITY
DOCUMENTS
WEBSITE

Please login for access to the knowledgebase. If you do not have a username, please register.

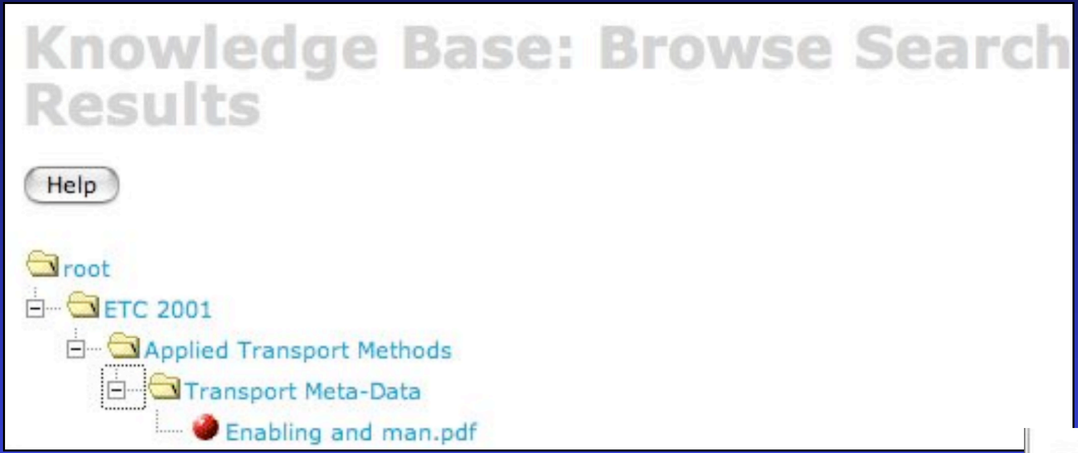
AET
Association for European Transport

The proceedings of the European Transport Conference 2001-2007

This AET KB is hosted on the N_KB_S framework first implemented for the R_KB.

Transport thesauri and automated keywords

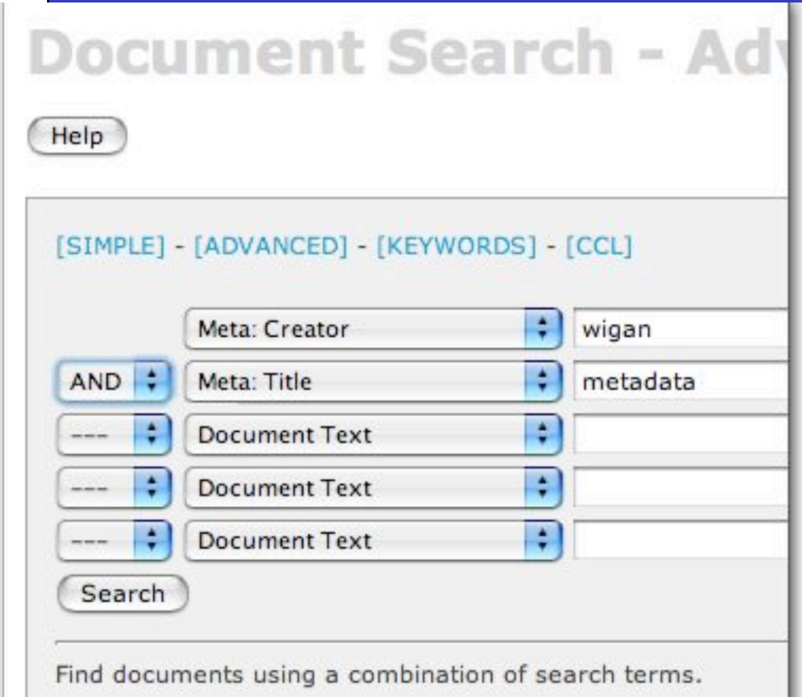
Metadata based Search and display



Knowledge Base: Browse Search Results

Find by 'Folder' navigation

Find by Boolean text and metadata Search



Document Search - Adv

Meta: Creator wigan
AND Meta: Title metadata
--- Document Text
--- Document Text
--- Document Text

Find documents using a combination of search terms.

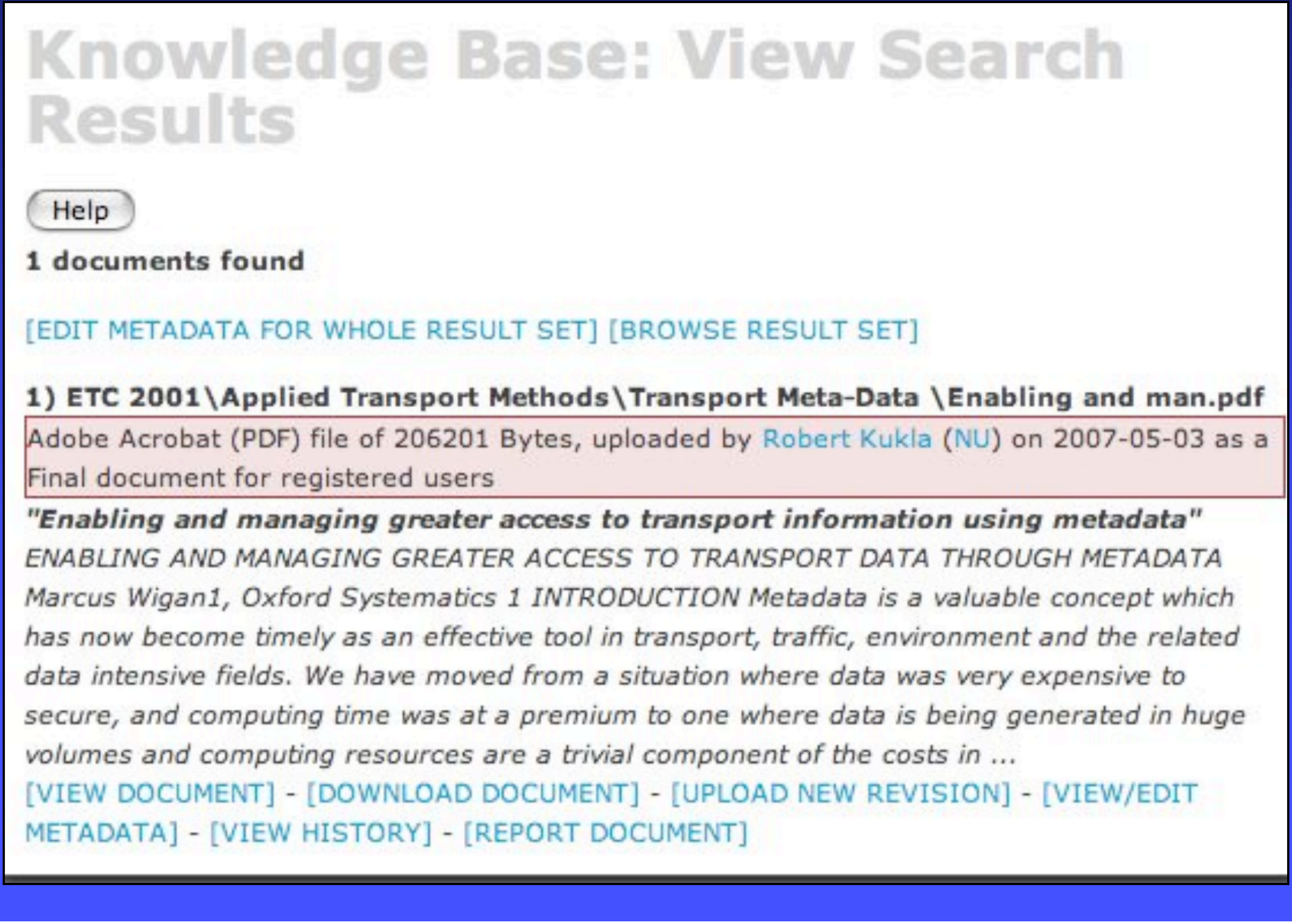
Oxford Systematics Australia

Transport thesauri and automated keywords

Quality of search limited by absence of metadata elements

- So must automate Keyword generation and input
- Use an English and a US spelling Thesaurus
  - ATRD from Australia (published Dec 2007)
  - NTDL from the US (also very recent)
- Match document free text with these Thesauri
- Remove single ocurrences and universal ones
- Use the resulting word lists to match to each document’s free text.
- Then inject matches as keywords for the document

Result of the 'advanced' Boolean search shown for AET



Knowledge Base: View Search Results

1 documents found

[EDIT METADATA FOR WHOLE RESULT SET] [BROWSE RESULT SET]

1) ETC 2001\Applied Transport Methods\Transport Meta-Data \Enabling and man.pdf

Adobe Acrobat (PDF) file of 206201 Bytes, uploaded by Robert Kukla (NU) on 2007-05-03 as a Final document for registered users

"Enabling and managing greater access to transport information using metadata"

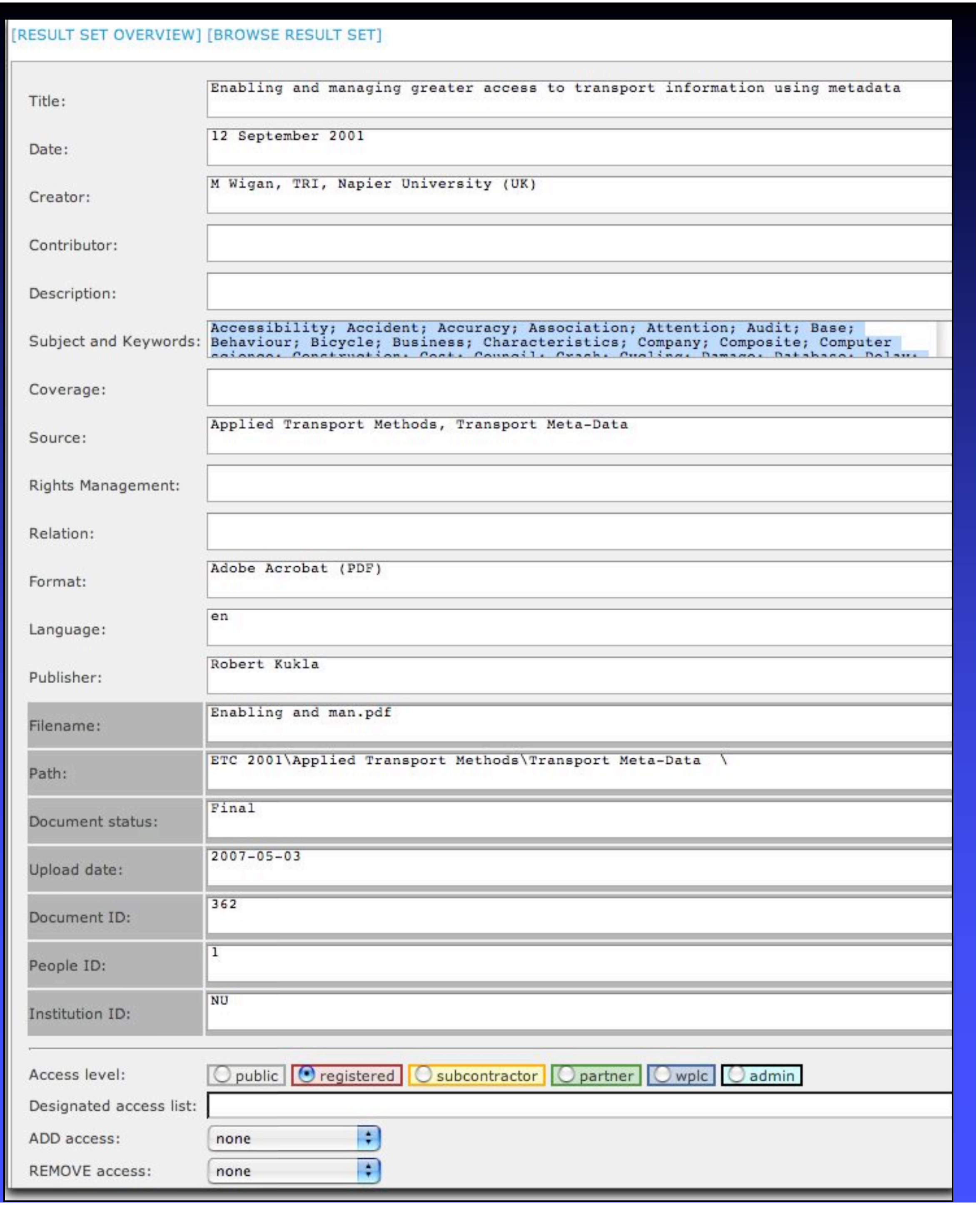
ENABLING AND MANAGING GREATER ACCESS TO TRANSPORT DATA THROUGH METADATA Marcus Wigan1, Oxford Systematics 1 INTRODUCTION Metadata is a valuable concept which has now become timely as an effective tool in transport, traffic, environment and the related data intensive fields. We have moved from a situation where data was very expensive to secure, and computing time was at a premium to one where data is being generated in huge volumes and computing resources are a trivial component of the costs in ...

[VIEW DOCUMENT] - [DOWNLOAD DOCUMENT] - [UPLOAD NEW REVISION] - [VIEW/EDIT METADATA] - [VIEW HISTORY] - [REPORT DOCUMENT]

Transport thesauri and automated keywords

There are now a large number of keywords in the scroll bar window

How does the “Edit Metadata” page look like now?

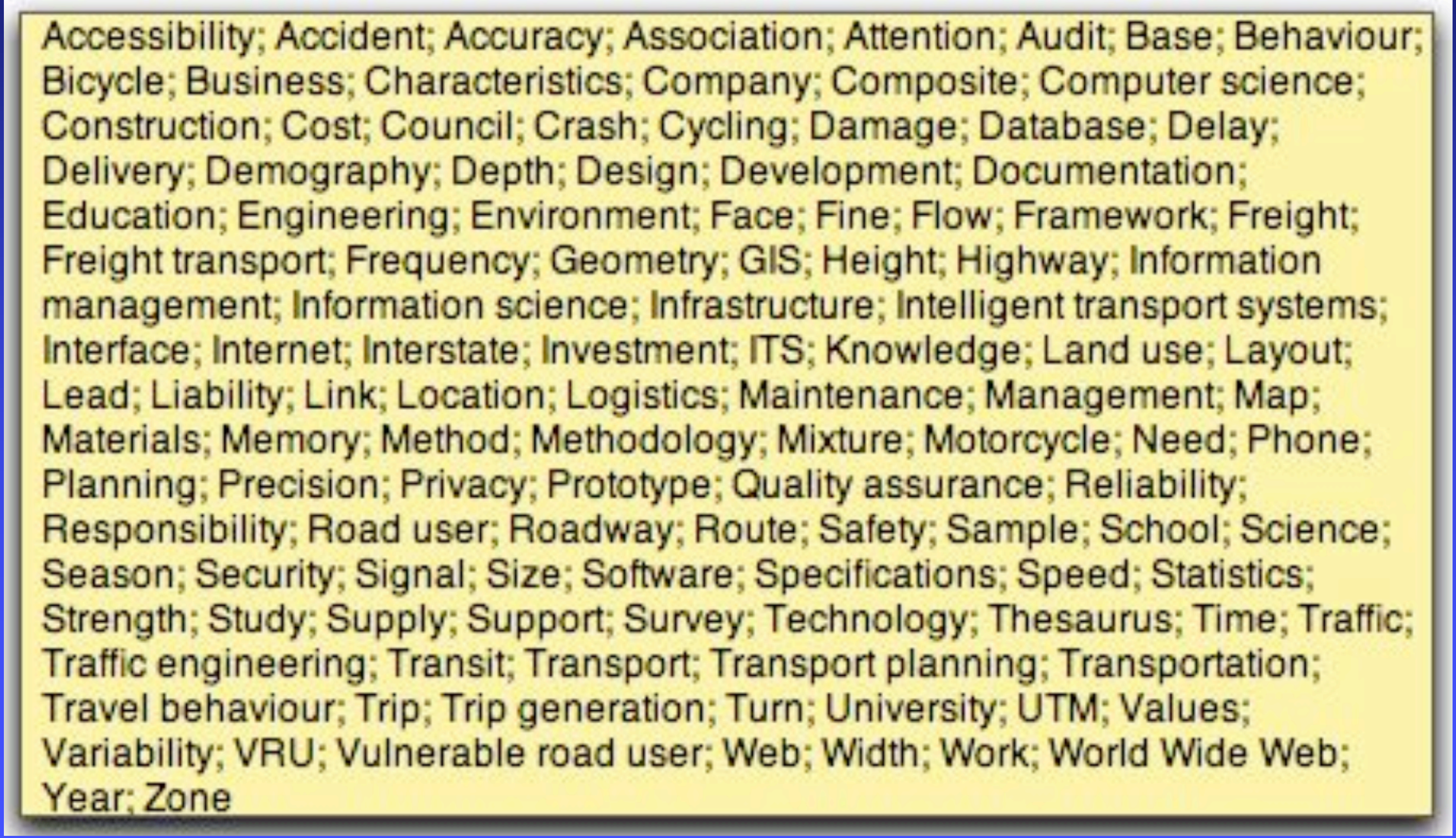


Title:	Enabling and managing greater access to transport information using metadata
Date:	12 September 2001
Creator:	M Wigan, TRI, Napier University (UK)
Contributor:	
Description:	
Subject and Keywords:	Accessibility; Accident; Accuracy; Association; Attention; Audit; Base; Behaviour; Bicycle; Business; Characteristics; Company; Composite; Computer science; Construction; Cost; Council; Crash; Cycling; Damage; Database; Delay;
Coverage:	
Source:	Applied Transport Methods, Transport Meta-Data
Rights Management:	
Relation:	
Format:	Adobe Acrobat (PDF)
Language:	en
Publisher:	Robert Kukla
Filename:	Enabling and man.pdf
Path:	ETC 2001\Applied Transport Methods\Transport Meta-Data \
Document status:	Final
Upload date:	2007-05-03
Document ID:	362
People ID:	1
Institution ID:	NU
Access level:	○ public ● registered ○ subcontractor ○ partner ○ wplc ○ admin
Designated access list:	
ADD access:	none
REMOVE access:	none

Oxford Systematics Australia

Transport thesauri and automated keywords

The automatically generated Keywords for this document



Accessibility; Accident; Accuracy; Association; Attention; Audit; Base; Behaviour; Bicycle; Business; Characteristics; Company; Composite; Computer science; Construction; Cost; Council; Crash; Cycling; Damage; Database; Delay; Delivery; Demography; Depth; Design; Development; Documentation; Education; Engineering; Environment; Face; Fine; Flow; Framework; Freight; Freight transport; Frequency; Geometry; GIS; Height; Highway; Information management; Information science; Infrastructure; Intelligent transport systems; Interface; Internet; Interstate; Investment; ITS; Knowledge; Land use; Layout; Lead; Liability; Link; Location; Logistics; Maintenance; Management; Map; Materials; Memory; Method; Methodology; Mixture; Motorcycle; Need; Phone; Planning; Precision; Privacy; Prototype; Quality assurance; Reliability; Responsibility; Road user; Roadway; Route; Safety; Sample; School; Science; Season; Security; Signal; Size; Software; Specifications; Speed; Statistics; Strength; Study; Supply; Support; Survey; Technology; Thesaurus; Time; Traffic; Traffic engineering; Transit; Transport; Transport planning; Transportation; Travel behaviour; Trip; Trip generation; Turn; University; UTM; Values; Variability; VRU; Vulnerable road user; Web; Width; Work; World Wide Web; Year; Zone

Commentary

- Searches using only Keywords now give good matches to free text searches
- Methodology papers get few Keywords
- This is a well known deficiency in transport Thesauri
- By using this technique methodology iteratively, keywords could be developed that work well in Thesauri and searches
- The use of an automated keyword field is clearly worthwhile
- Searches can be Boolean limited by combining the Keyword Metadata date field in conjunction with others