Experiences in Deploying Public Metadata Analysis Tools

David Nichols

Department of Computer Science

University of Waikato

Hamilton, New Zealand

KO TE TANGATA

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Outline

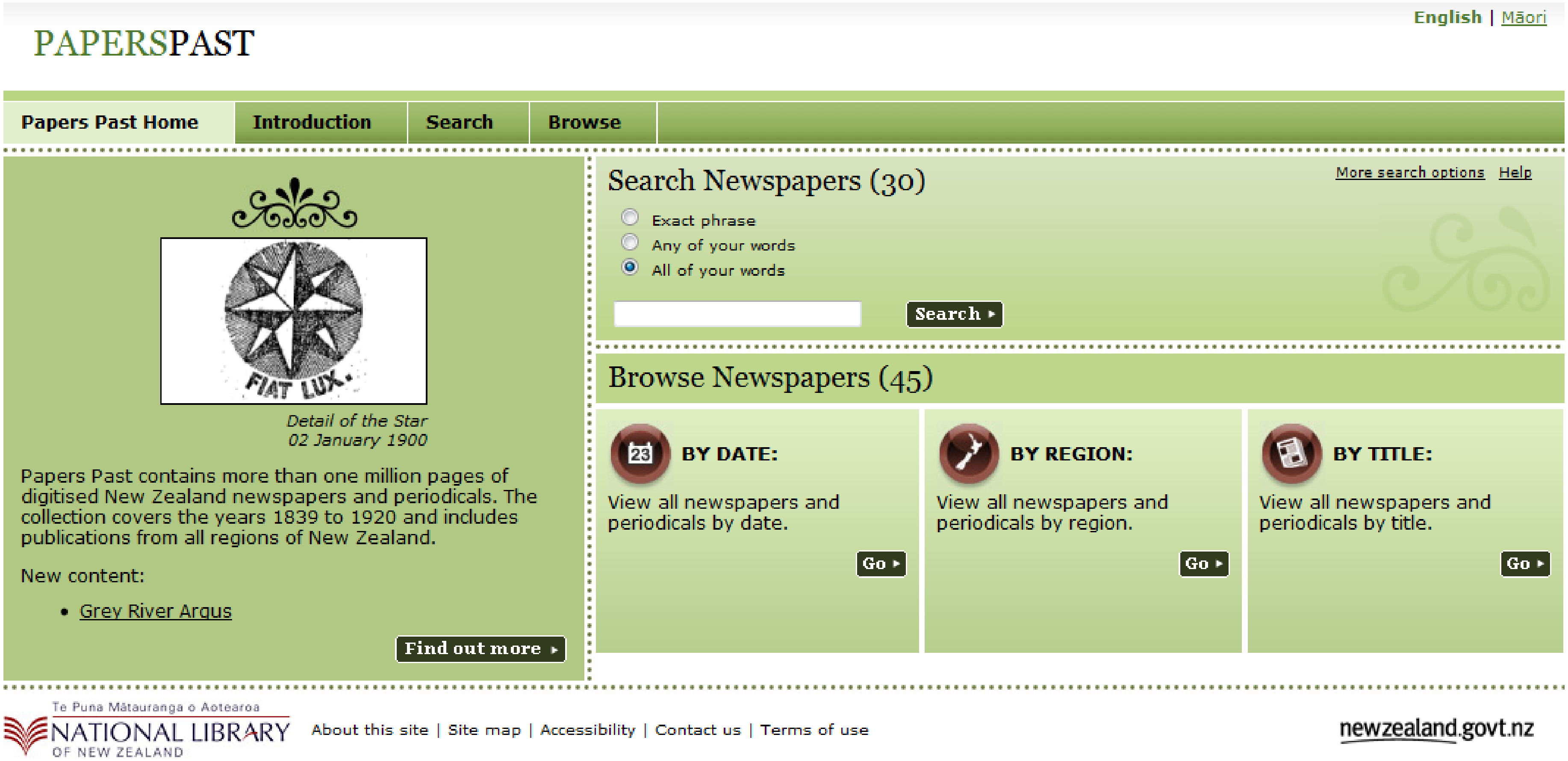
- Motivation & Background
- The Mat web tool
- Demo / Examples
- Qualitative feedback
- New beta features
- The KRIS web tool
- Demo / Examples
- Summary
- *Postscript*

Original Aim

- Provide better feedback to collection creators in Greenstone 3

- Previously...

paperspast.natlib.govt.nz



PAPERSPAST

English | Māori

Papers Past Home | Introduction | Search | Browse

Detail of the Star
02 January 1900

Papers Past contains more than one million pages of digitised New Zealand newspapers and periodicals. The collection covers the years 1839 to 1920 and includes publications from all regions of New Zealand.

New content:

- Grey River Argus

Find out more ▸

Search Newspapers (30)

More search options Help

○ Exact phrase
○ Any of your words
● All of your words

Search ▸

Browse Newspapers (45)

BY DATE: View all newspapers and periodicals by date. Go ▸

BY REGION: View all newspapers and periodicals by region. Go ▸

BY TITLE: View all newspapers and periodicals by title. Go ▸

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

About this site | Site map | Accessibility | Contact us | Terms of use

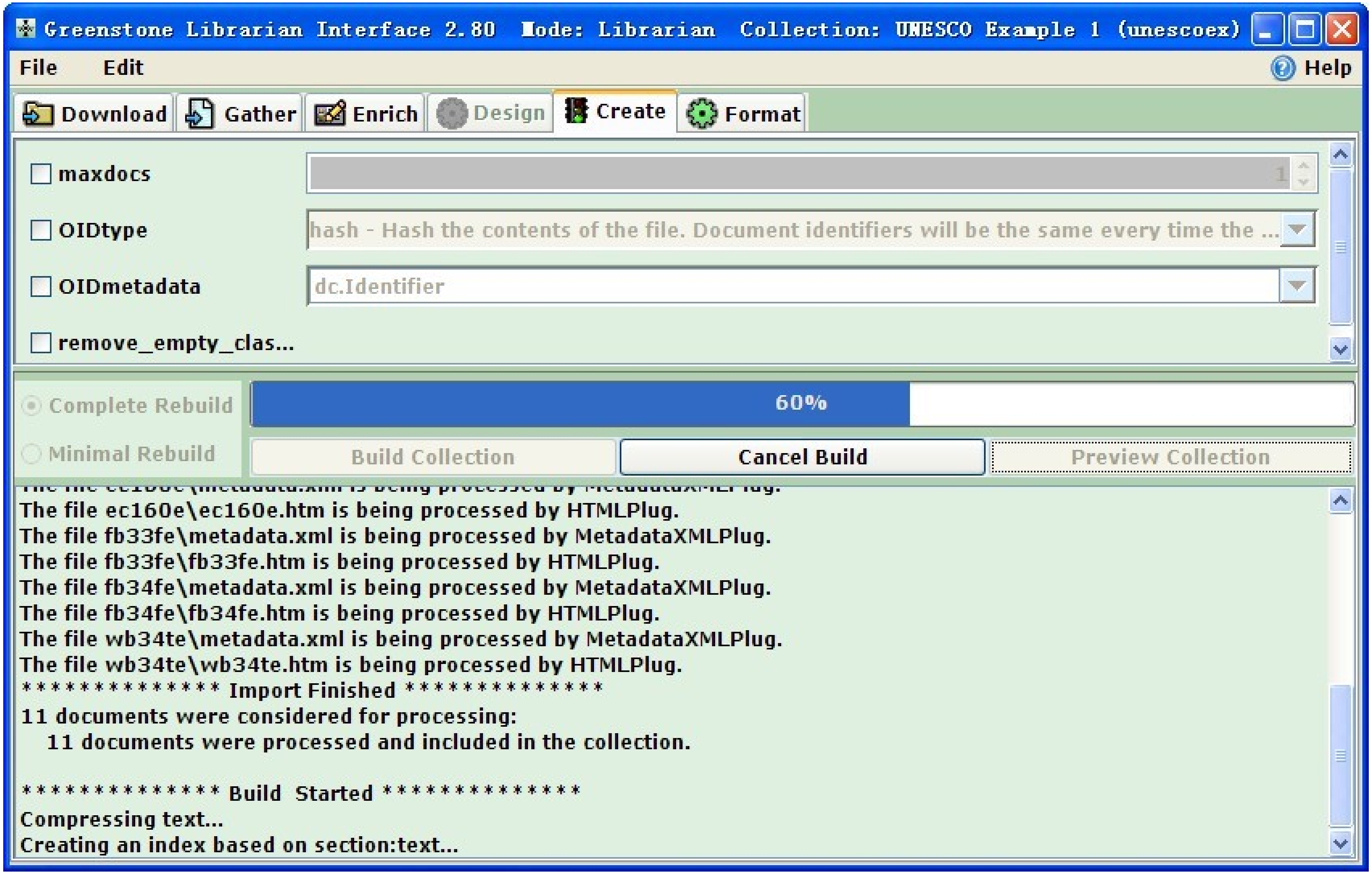
newzealand.govt.nz

1.2 + million newspaper pages
600,000 + searchable

OCR
METS/ALTO

dlconsulting.com

Creation/Workflow interface



□ maxdocs 1

□ OIDtype hash - Hash the contents of the file. Document identifiers will be the same every time the ...

□ OIDmetadata dc.Identifier

□ remove_empty_clas...

○ Complete Rebuild

○ Minimal Rebuild

60%

```
The file ec160e\metadata.xml is being processed by MetadataXMLPlug.
The file ec160e\ec160e.htm is being processed by HTMLPlug.
The file fb33fe\metadata.xml is being processed by MetadataXMLPlug.
The file fb33fe\fb33fe.htm is being processed by HTMLPlug.
The file fb34fe\metadata.xml is being processed by MetadataXMLPlug.
The file fb34fe\fb34fe.htm is being processed by HTMLPlug.
The file wb34te\metadata.xml is being processed by MetadataXMLPlug.
The file wb34te\wb34te.htm is being processed by HTMLPlug.
************** Import Finished **************
11 documents were considered for processing:
  11 documents were processed and included in the collection.

************** Build  Started **************
Compressing text...
Creating an index based on section:text...
```

Creator/Maintainer Feedback

- Current Feedback =
  - Running system +
  - build date +
  - number of documents processed +
  - Some low-level details about compression & weights
- difficult to manually identify metadata quality issues
- DL systems need *automated* tools
- Are there experience reports of existing systems?

Metadata Quality?

- metadata quality criteria
  - e.g. Bruce & Hillmann (2003)
- Beal (2005)
  - librarytypos.blogspot.com
- Several surveys (Shreeves, Ward, Efron etc) of DC element usage
  - Code not re-used?

✓ completeness

✓ accuracy

✓ provenance

✓ conformance to expectations

✓ logical consistency and coherence

✓ timeliness

✓ accessibility

"Throughout the eprints community there is an increasing awareness of the need for improvement in the quality of metadata and in associated quality assurance mechanisms"

Guy, Powell & Day (2004)

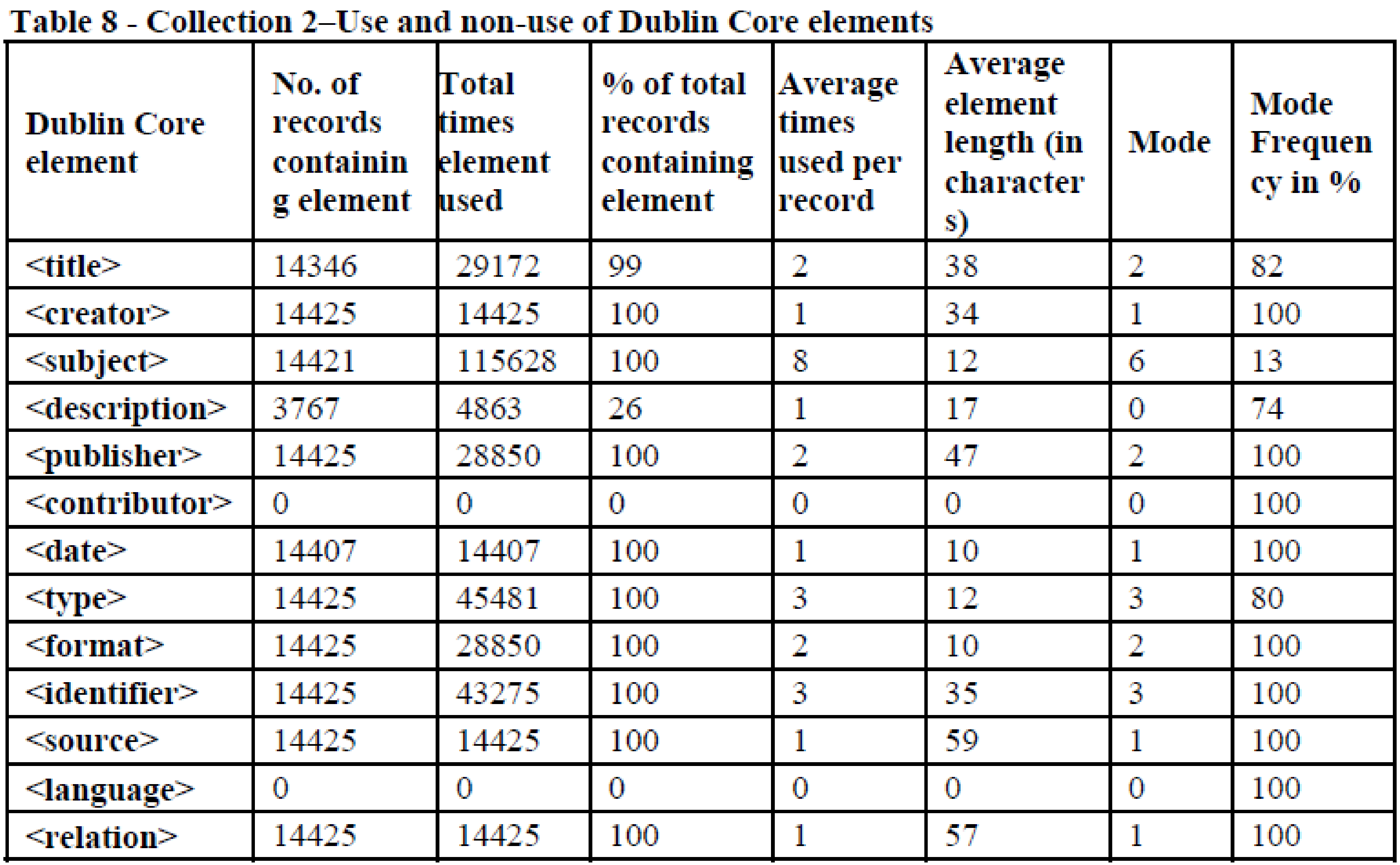


Table 8 - Collection 2–Use and non-use of Dublin Core elements

Dublin Core element	No. of records containing element	Total times element used	% of total records containing element	Average times used per record	Average element length (in characters)	Mode	Mode Frequency in %
<title>	14346	29172	99	2	38	2	82
<creator>	14425	14425	100	1	34	1	100
<subject>	14421	115628	100	8	12	6	13
<description>	3767	4863	26	1	17	0	74
<publisher>	14425	28850	100	2	47	2	100
<contributor>	0	0	0	0	0	0	100
<date>	14407	14407	100	1	10	1	100
<type>	14425	45481	100	3	12	3	80
<format>	14425	28850	100	2	10	2	100
<identifier>	14425	43275	100	3	35	3	100
<source>	14425	14425	100	1	59	1	100
<language>	0	0	0	0	0	0	100
<relation>	14425	14425	100	1	57	1	100

Shreeves *et al* (2005)

Metadata Visualisation

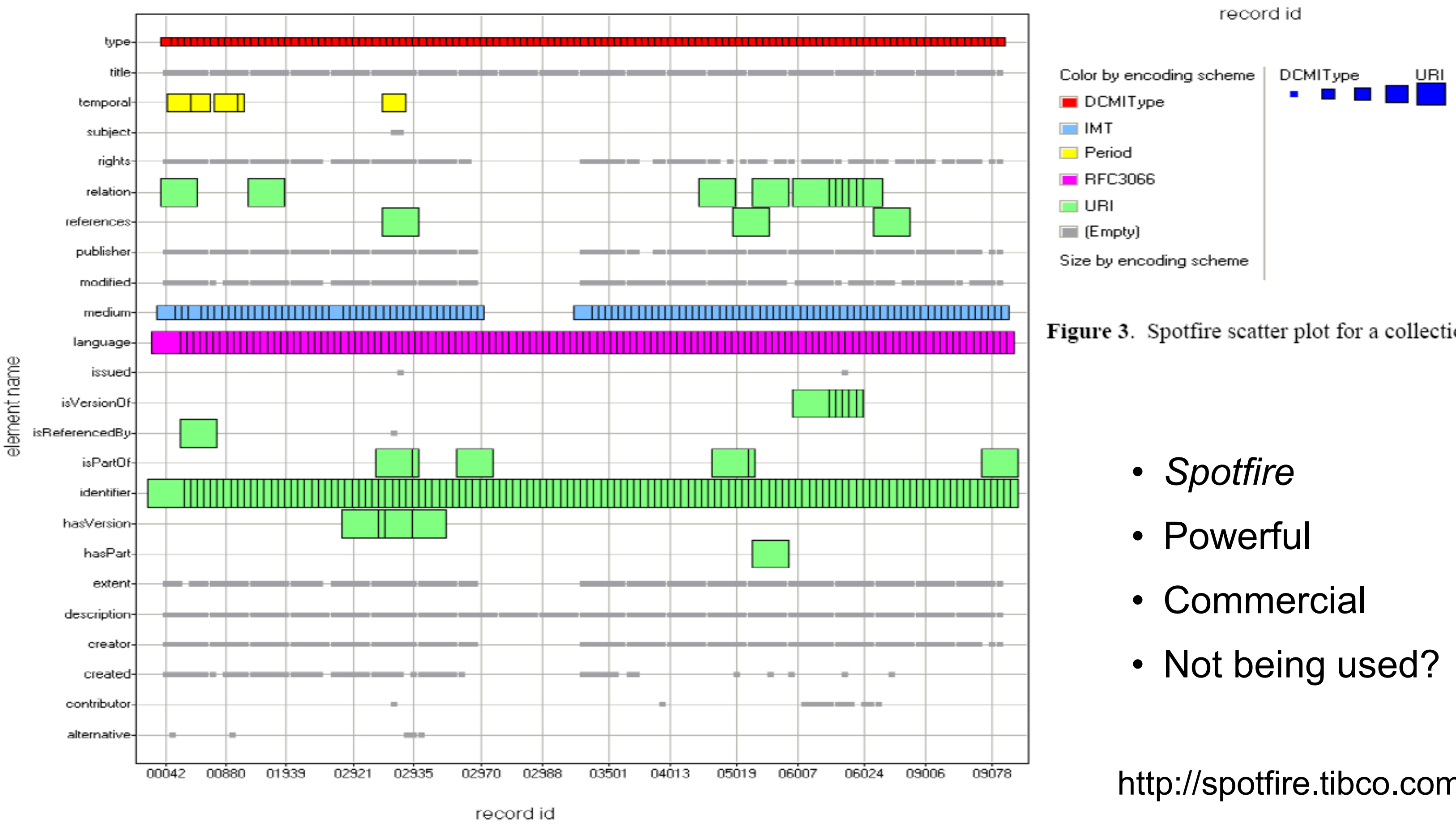


Figure 3. Spotfire scatter plot for a collection's metadata

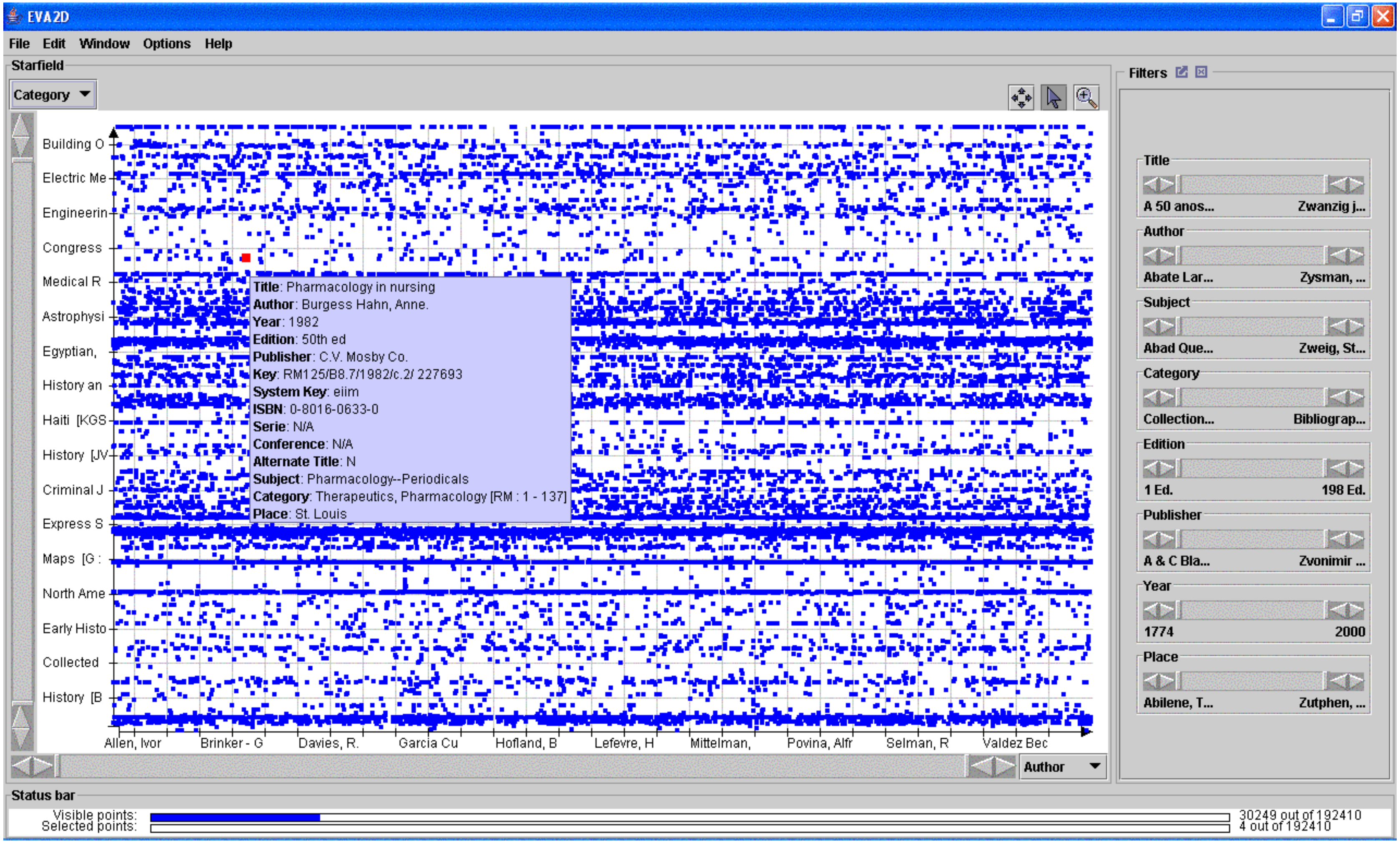
- *Spotfire*
- Powerful
- Commercial
- Not being used?

http://spotfire.tibco.com/

"the use of data visualization software can significantly improve efficiency and thoroughness of metadata evaluation"

Dushay and Hillmann (2003)

Starfield visualisation of library catalogue



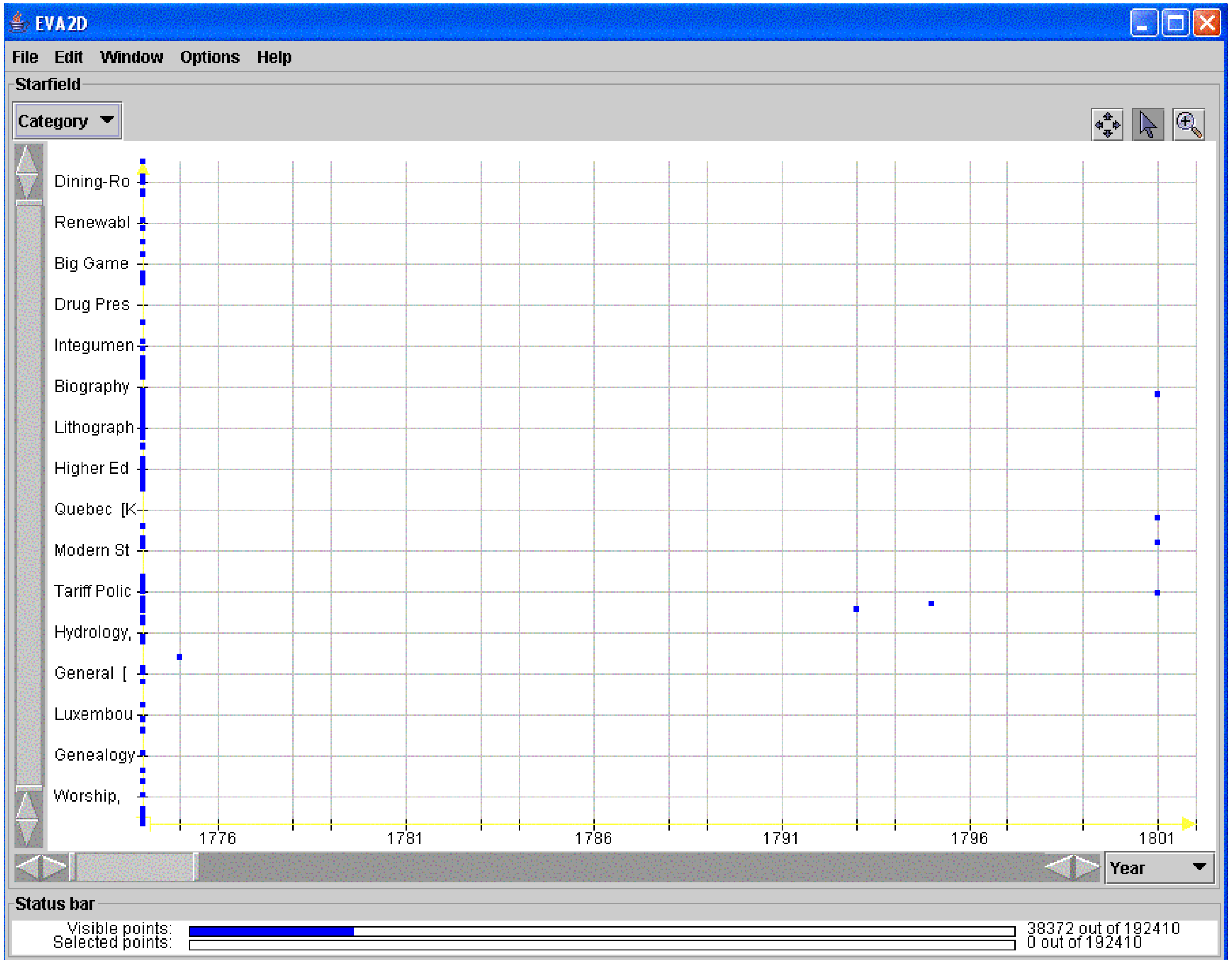
Title: Pharmacology in nursing
Author: Burgess Hahn, Anne.
Year: 1982
Edition: 50th ed
Publisher: C.V. Mosby Co.
Key: RM125/B8.7/1982/c.2/ 227693
System Key: eiim
ISBN: 0-8016-0633-0
Serie: N/A
Conference: N/A
Alternate Title: N
Subject: Pharmacology--Periodicals
Category: Therapeutics, Pharmacology [RM : 1 - 137]
Place: St. Louis

Visible points: 30249 out of 192410
Selected points: 4 out of 192410

190,000 OPAC records from UDLAP

Sánchez, Twidale, Nichols & Silva (2005)

Error and pattern detection

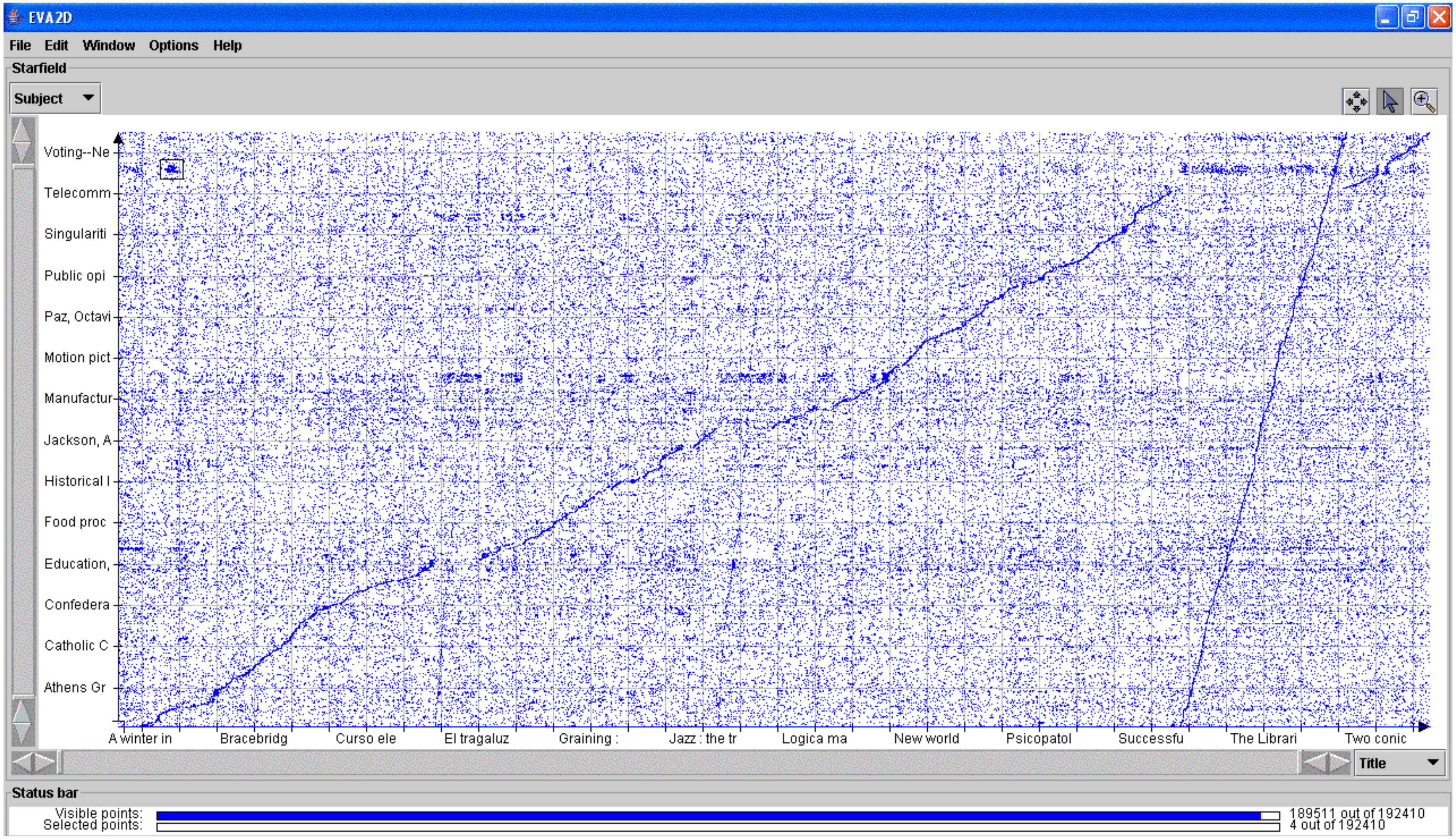


Visible points: 38372 out of 192410
Selected points: 0 out of 192410

- EVA2D
- UDLAP
- Mainly visual
- Java App

Quiz: what does this represent?

Subject



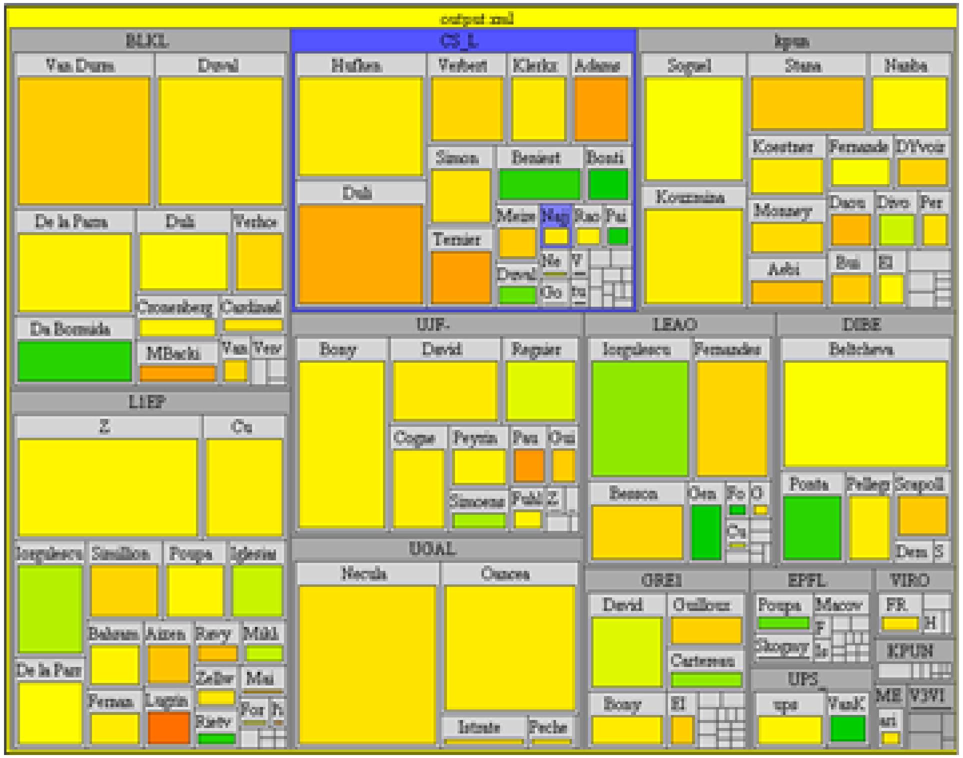
Visible points: 189511 out of 192410
Selected points: 4 out of 192410

Title

Uses of metadata quality information

- Error detection
- Collection understanding
- Input to other tools →
    - Quality mapped to colour
    - … on a *Treemap*

http://www.cs.umd.edu/hcil/treemap/



"Visualization of the Textual Information Content of the ARIADNE Repository"

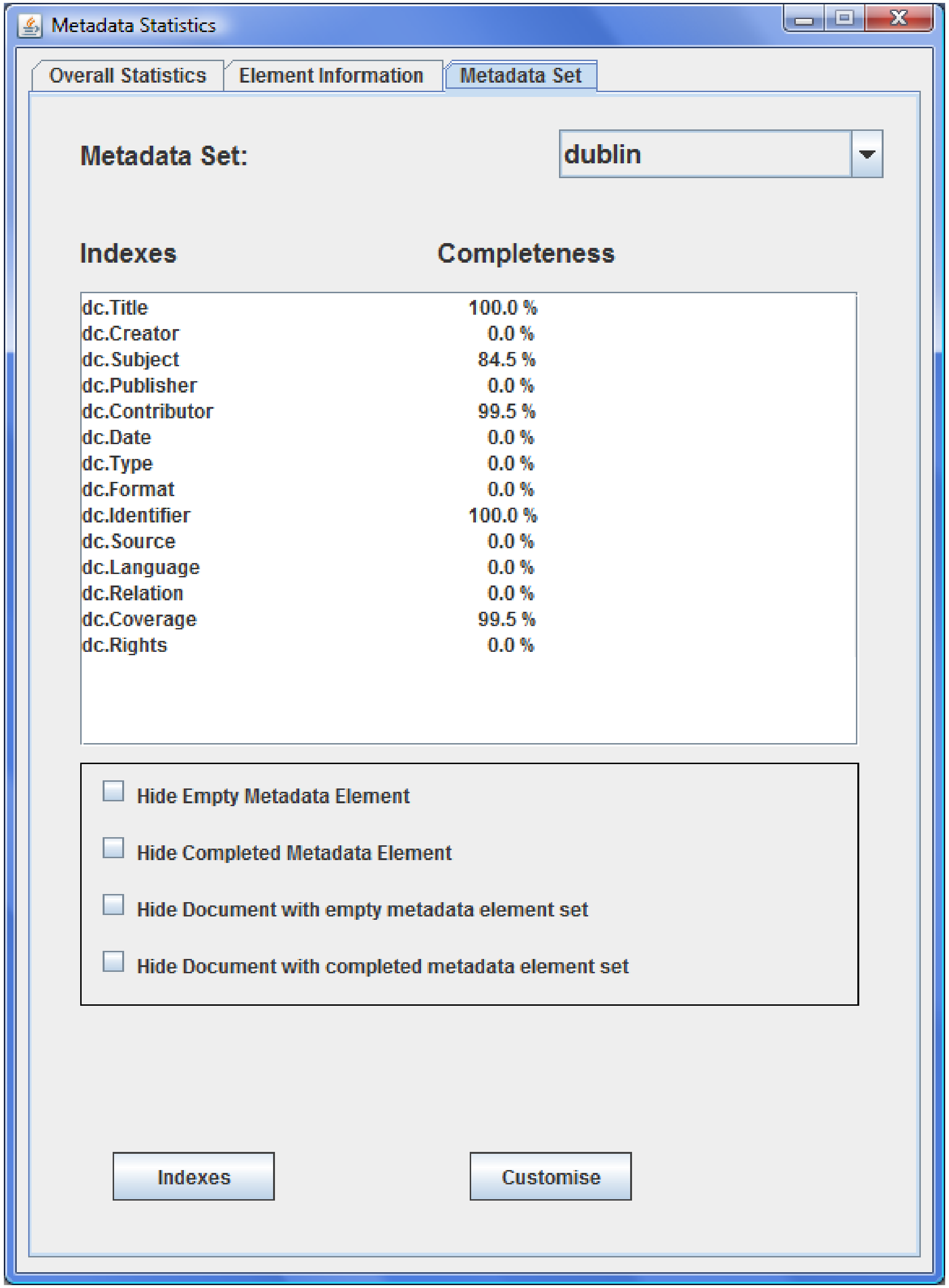
Ochoa and Duval (2006)

So ...

- We didn’t find anything we could just plug in to Greenstone
- Not all the quality metrics are easily computable, but at least...
- It would be good to have functionality similar to:
  - Statistics from the DC usage surveys
  - Visualisations

First prototype

- Java-based
- Analyses collections
- Generates statistics
    - Chosen based on guesswork and DC surveys
- Simple visualisation



Metadata Set: dublin

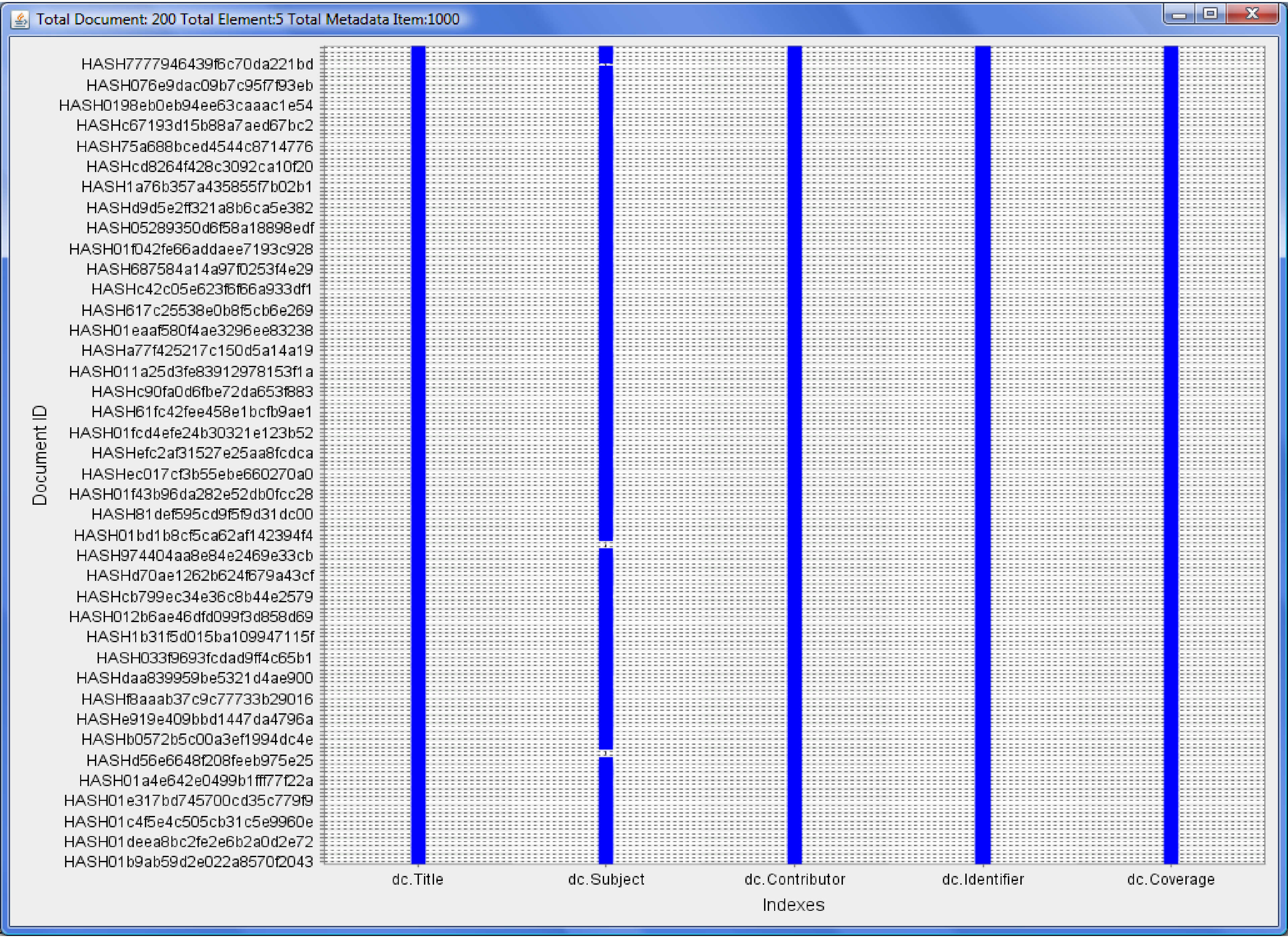
Indexes	Completeness
dc.Title	100.0 %
dc.Creator	0.0 %
dc.Subject	84.5 %
dc.Publisher	0.0 %
dc.Contributor	99.5 %
dc.Date	0.0 %
dc.Type	0.0 %
dc.Format	0.0 %
dc.Identifier	100.0 %
dc.Source	0.0 %
dc.Language	0.0 %
dc.Relation	0.0 %
dc.Coverage	99.5 %
dc.Rights	0.0 %

☐ Hide Empty Metadata Element

☐ Hide Completed Metadata Element

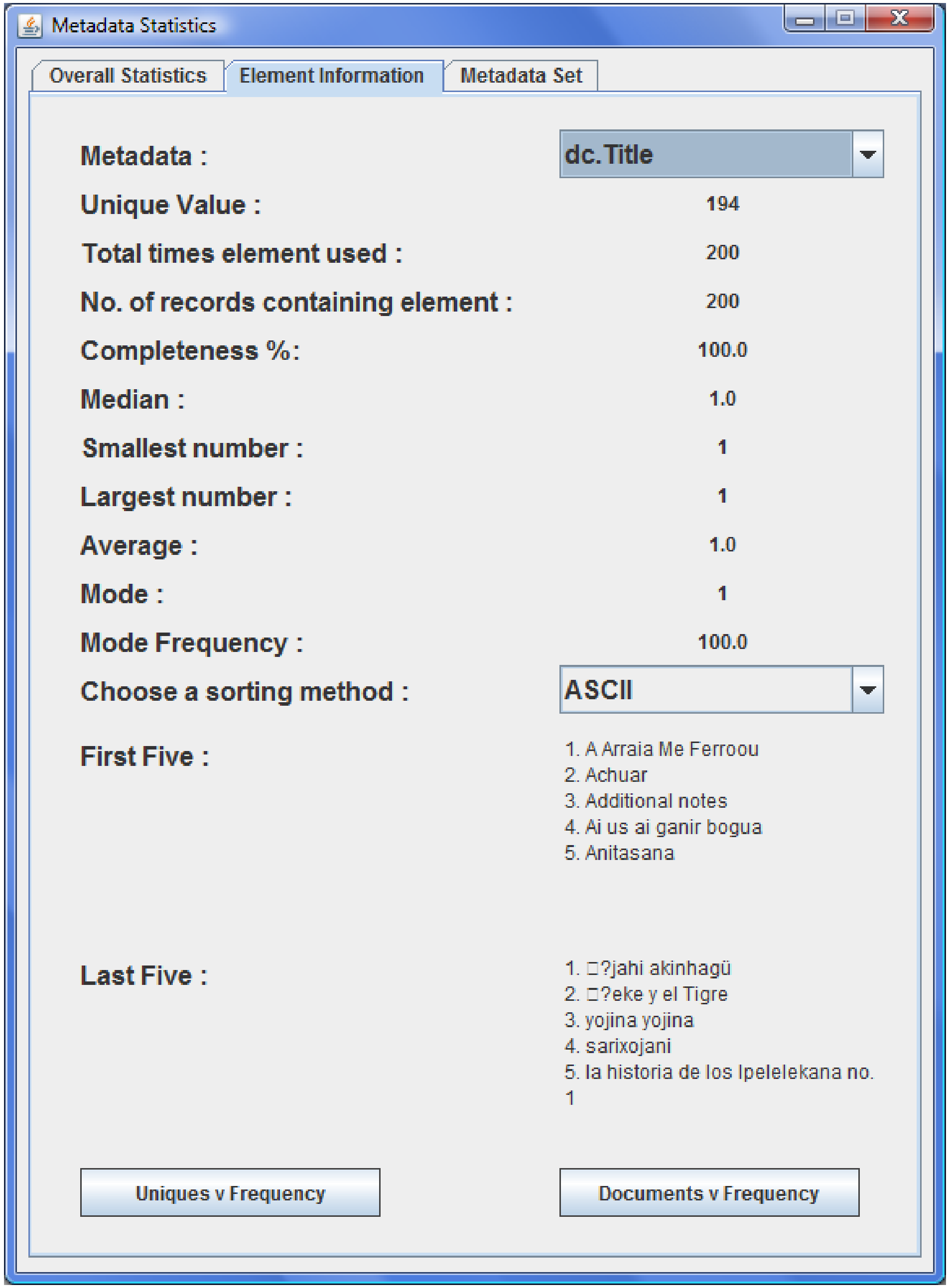
☐ Hide Document with empty metadata element set

☐ Hide Document with completed metadata element set



Total Document: 200 Total Element:5 Total Metadata Item:1000

Document ID
HASH7777946439f6c70da221bd
HASH076e9dac09b7c95f7f93eb
HASH0198eb0eb94ee63caaac1e54
HASHc67193d15b88a7aed67bc2
HASH75a688bced4544c8714776
HASHcd8264f428c3092ca10f20
HASH1a76b357a435855f7b02b1
HASHd9d5e2ff321a8b6ca5e382
HASH05289350d6f58a18898edf
HASH01f042fe66addaee7193c928
HASH687584a14a97f0253f4e29
HASHc42c05e623f6f66a933df1
HASH617c25538e0b8f5cb6e269
HASH01eaaf580f4ae3296ee83238
HASHa77f425217c150d5a14a19
HASH011a25d3fe83912978153f1a
HASHc90fa0d6fbe72da653f883
HASH61fc42fee458e1bcfb9ae1
HASH01fcd4efe24b30321e123b52
HASHefc2af31527e25aa8fcdca
HASHec017cf3b55ebe660270a0
HASH01f43b96da282e52db0fcc28
HASH81def595cd9f5f9d31dc00
HASH01bd1b8cf5ca62af142394f4
HASH974404aa8e84e2469e33cb
HASHd70ae1262b624f679a43cf
HASHcb799ec34e36c8b44e2579
HASH012b6ae46dfd099f3d858d69
HASH1b31f5d015ba109947115f
HASH033f9693fcdad9ff4c65b1
HASHdaa839959be5321d4ae900
HASHf8aaab37c9c77733b29016
HASHe919e409bbd1447da4796a
HASHb0572b5c00a3ef1994dc4e
HASHd56e6648f208feeb975e25
HASH01a4e642e0499b1fff77f22a
HASH01e317bd745700cd35c779f9
HASH01c4f5e4c505cb31c5e9960e
HASH01deea8bc2fe2e6b2a0d2e72
HASH01b9ab59d2e022a8570f2043
dc.Title
dc.Subject
dc.Contributor
dc.Identifier
dc.Coverage
Indexes



Metadata : dc.Title

Unique Value : 194

Total times element used : 200

No. of records containing element : 200

Completeness %: 100.0

Median : 1.0

Smallest number : 1

Largest number : 1

Average : 1.0

Mode : 1

Mode Frequency : 100.0

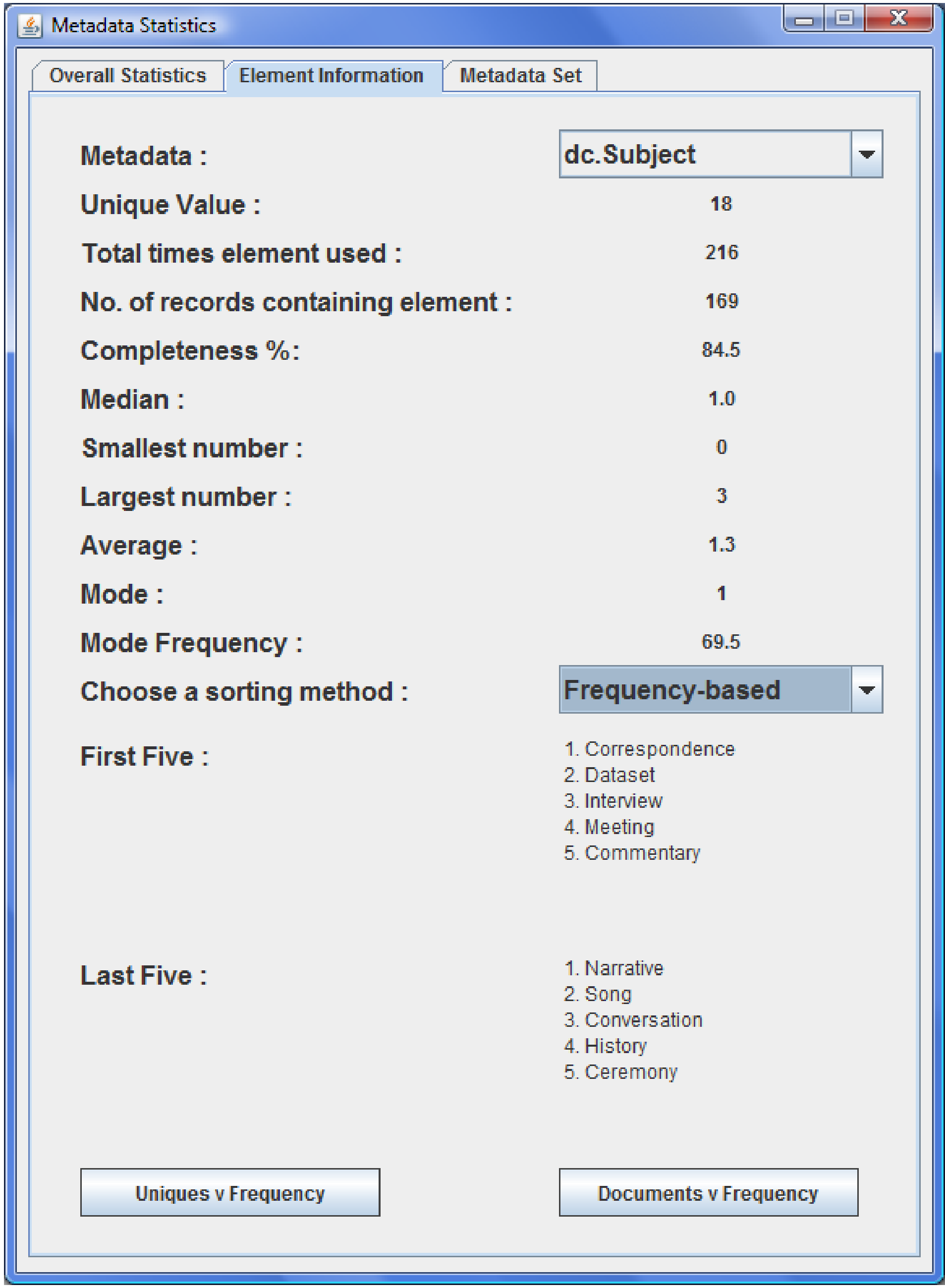
Choose a sorting method : ASCII

First Five :

1. A Arraia Me Ferroou
2. Achuar
3. Additional notes
4. Ai us ai ganir bogua
5. Anitasana

Last Five :

1. □?jahi akinhagü
2. □?eke y el Tigre
3. yojina yojina
4. sarixojani
5. la historia de los Ipelelekana no. 1



Metadata : dc.Subject

Unique Value : 18

Total times element used : 216

No. of records containing element : 169

Completeness %: 84.5

Median : 1.0

Smallest number : 0

Largest number : 3

Average : 1.3

Mode : 1

Mode Frequency : 69.5

Choose a sorting method : Frequency-based

First Five :

1. Correspondence
2. Dataset
3. Interview
4. Meeting
5. Commentary

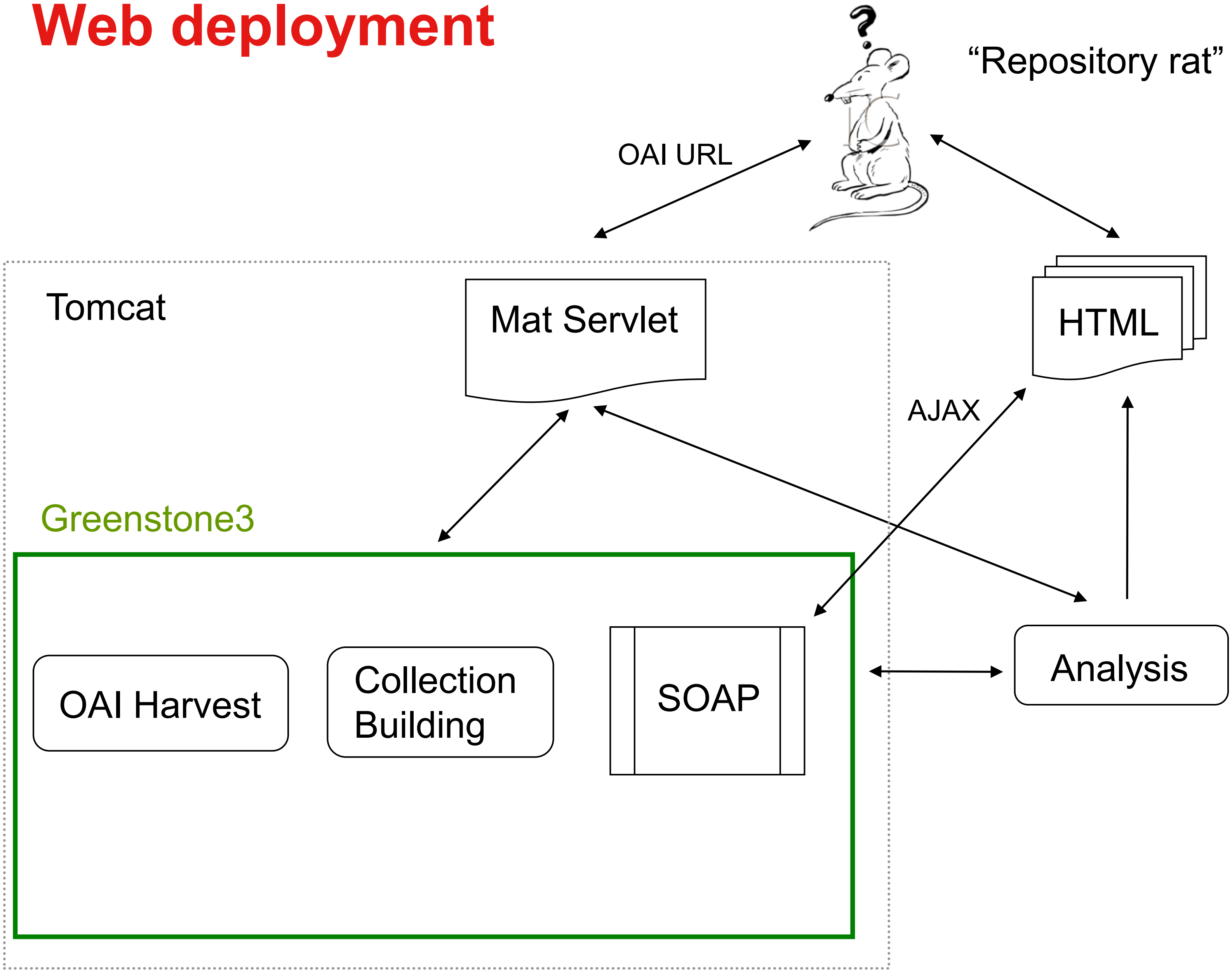
Last Five :

1. Narrative
2. Song
3. Conversation
4. History
5. Ceremony

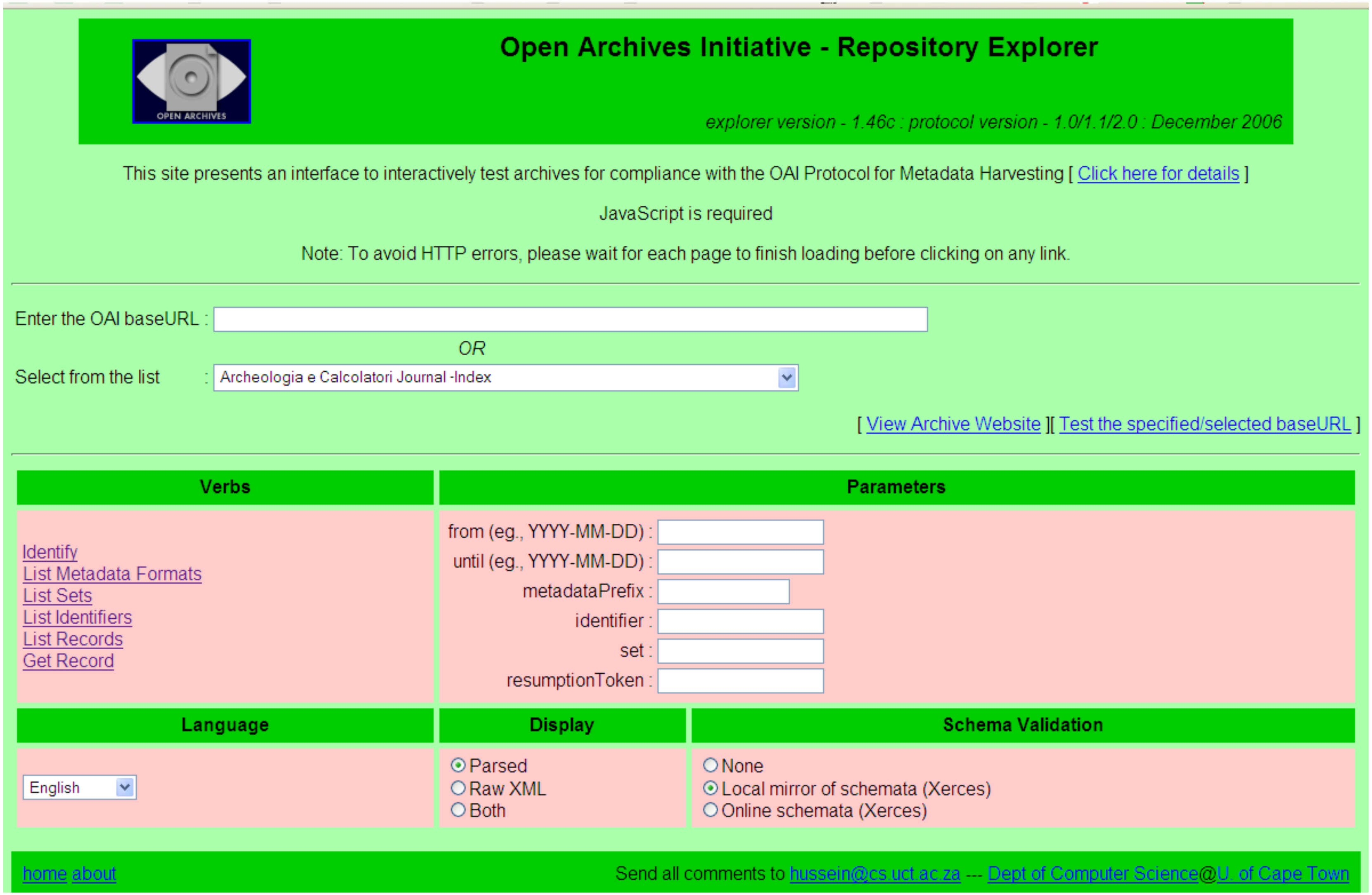
Revised aim

- Understand *requirements* for metadata quality tool
- Migrated Java prototype → Web
  - Simplicity of use
  - Wide use and (hopefully) feedback
    - esp. Institutional Repository community
- We built a web-based Metadata Analysis Tool

Web deployment



OAI Repository Explorer @ UCT



Open Archives Initiative - Repository Explorer

explorer version - 1.46c : protocol version - 1.0/1.1/2.0 : December 2006

This site presents an interface to interactively test archives for compliance with the OAI Protocol for Metadata Harvesting [Click here for details]

JavaScript is required

Note: To avoid HTTP errors, please wait for each page to finish loading before clicking on any link.

Enter the OAI baseURL :

OR

Select from the list : Archeologia e Calcolatori Journal -Index

[View Archive Website][Test the specified/selected baseURL]

Verbs	Parameters
Identify List Metadata Formats List Sets List Identifiers List Records Get Record	from (eg., YYYY-MM-DD) : until (eg., YYYY-MM-DD) : metadataPrefix : identifier : set : resumptionToken :

Language	Display	Schema Validation
English	● Parsed ○ Raw XML ○ Both	○ None ● Local mirror of schemata (Xerces) ○ Online schemata (Xerces)

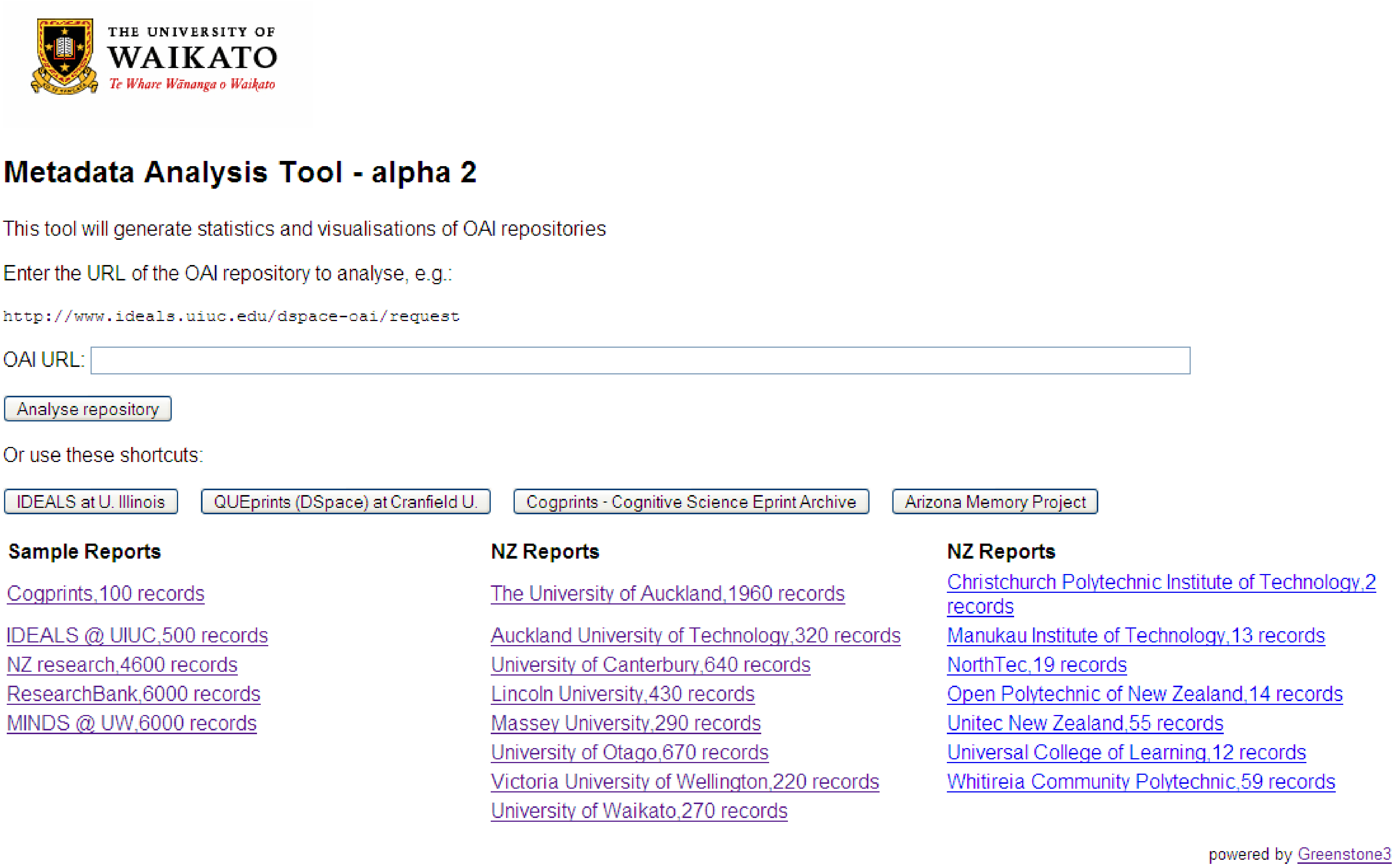
home about

Send all comments to hussein@cs.uct.ac.za --- Dept of Computer Science@U. of Cape Town

http://re.cs.uct.ac.za/

Suleman (JCDL '01)

http://www.nzdl.org/greenstone3/mat


THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato
Analyse repository
IDEALS at U. Illinois
QUEprints (DSpace) at Cranfield U.
Cogprints - Cognitive Science Eprint Archive
Arizona Memory Project

Metadata Analysis Tool - alpha 2

This tool will generate statistics and visualisations of OAI repositories

Enter the URL of the OAI repository to analyse, e.g.:

```
http://www.ideals.uiuc.edu/dspace-oai/request
```

OAI URL:

Or use these shortcuts:

Sample Reports

- Cogprints,100 records
- IDEALS @ UIUC,500 records
- NZ research,4600 records
- ResearchBank,6000 records
- MINDS @ UW,6000 records

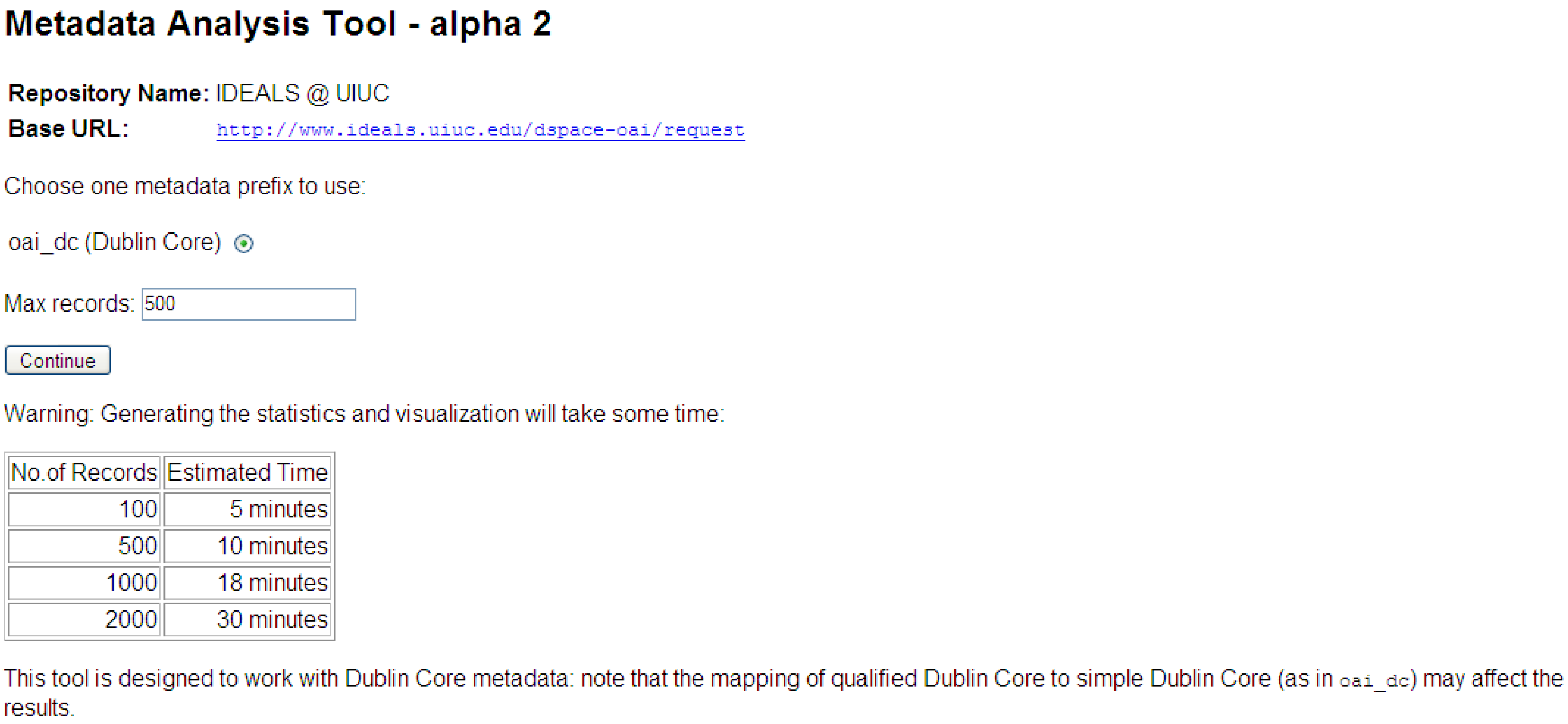
NZ Reports

- The University of Auckland,1960 records
- Auckland University of Technology,320 records
- University of Canterbury,640 records
- Lincoln University,430 records
- Massey University,290 records
- University of Otago,670 records
- Victoria University of Wellington,220 records
- University of Waikato,270 records

NZ Reports

- Christchurch Polytechnic Institute of Technology,2 records
- Manukau Institute of Technology,13 records
- NorthTec,19 records
- Open Polytechnic of New Zealand,14 records
- Unitec New Zealand,55 records
- Universal College of Learning,12 records
- Whitireia Community Polytechnic,59 records

powered by Greenstone3



Metadata Analysis Tool - alpha 2

Repository Name: IDEALS @ UIUC
Base URL: http://www.ideals.uiuc.edu/dspace-oai/request

Choose one metadata prefix to use:

oai_dc (Dublin Core) ●

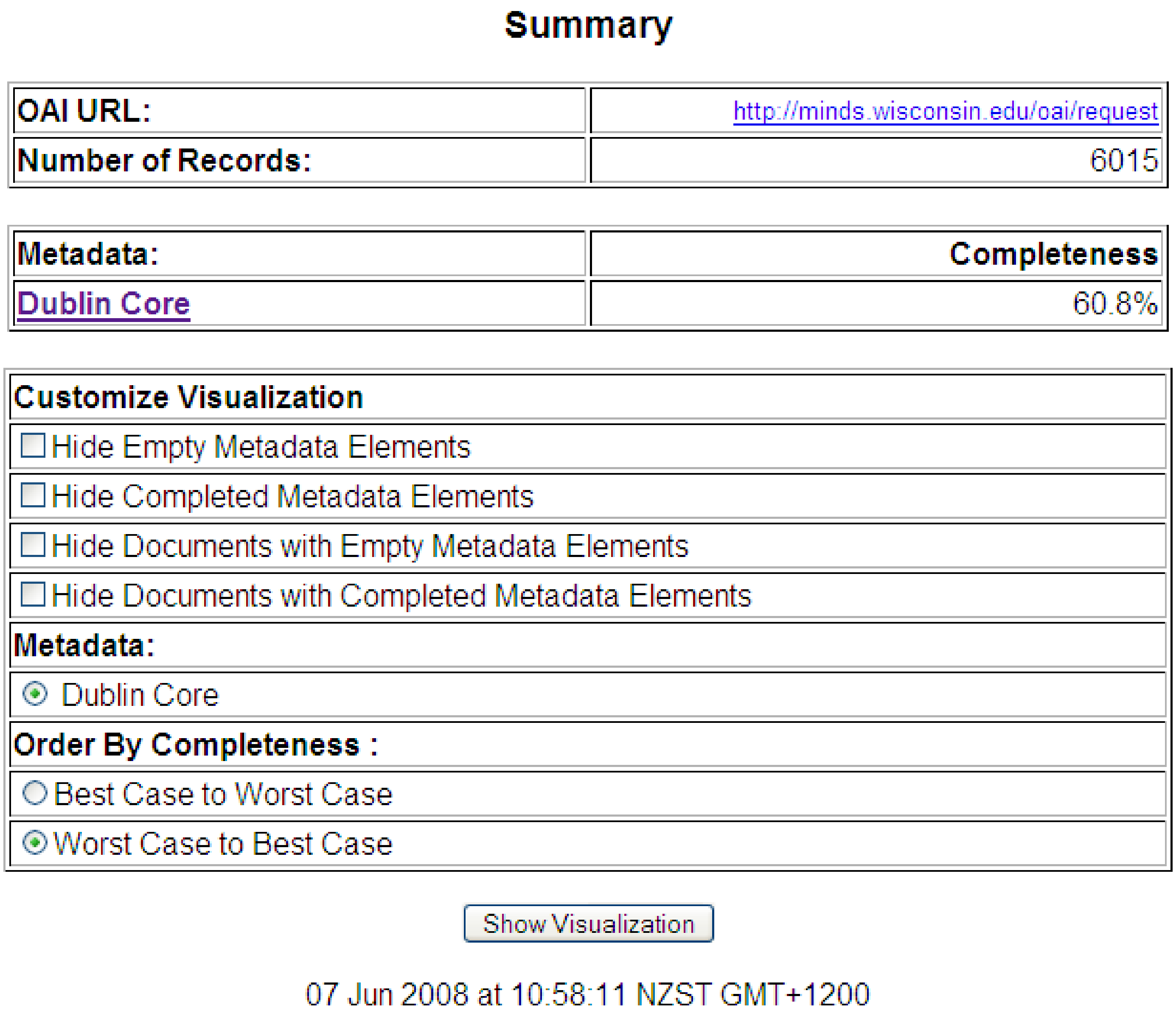
Max records: 500

Warning: Generating the statistics and visualization will take some time:

No.of Records	Estimated Time
100	5 minutes
500	10 minutes
1000	18 minutes
2000	30 minutes

This tool is designed to work with Dublin Core metadata: note that the mapping of qualified Dublin Core to simple Dublin Core (as in `oai_dc`) may affect the results.

Simply the results of an *Identify* request



Summary

OAI URL:	http://minds.wisconsin.edu/oai/request
Number of Records:	6015

Metadata:	Completeness
Dublin Core	60.8%

Customize Visualization

☐ Hide Empty Metadata Elements

☐ Hide Completed Metadata Elements

☐ Hide Documents with Empty Metadata Elements

☐ Hide Documents with Completed Metadata Elements

Metadata:

● Dublin Core

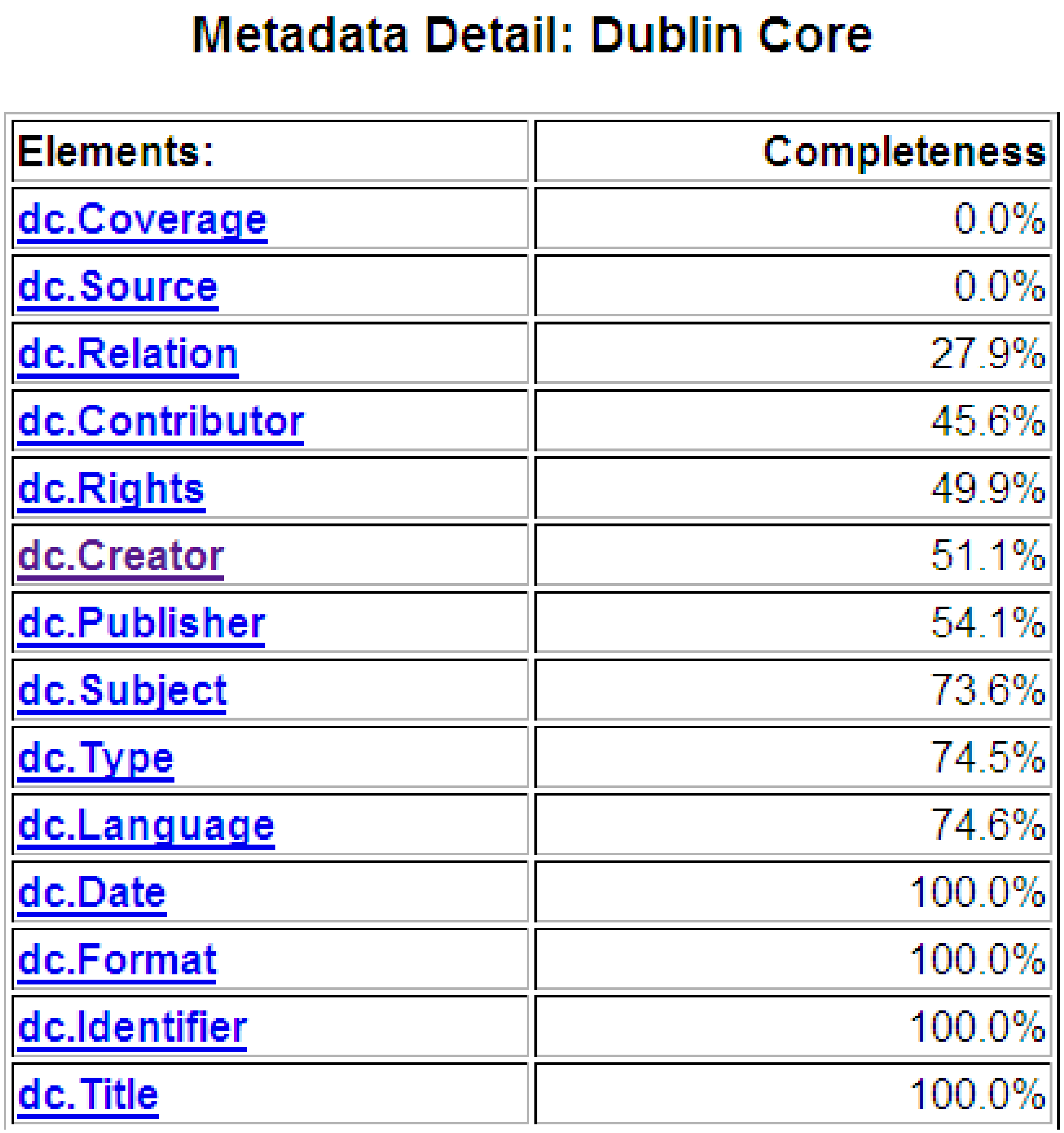
Order By Completeness :

○ Best Case to Worst Case

● Worst Case to Best Case

07 Jun 2008 at 10:58:11 NZST GMT+1200

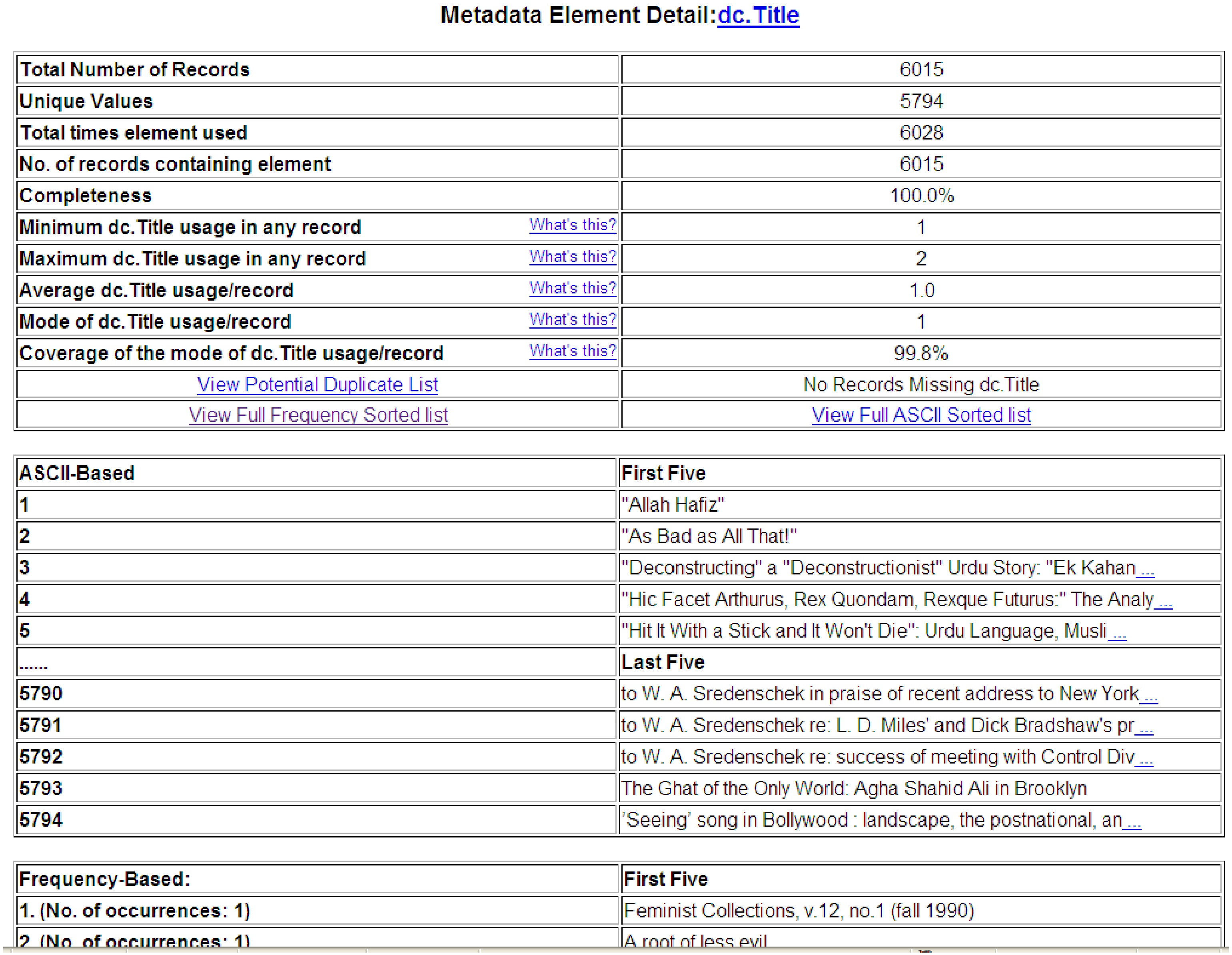
Overview



Metadata Detail: Dublin Core

Elements:	Completeness
dc.Coverage	0.0%
dc.Source	0.0%
dc.Relation	27.9%
dc.Contributor	45.6%
dc.Rights	49.9%
dc.Creator	51.1%
dc.Publisher	54.1%
dc.Subject	73.6%
dc.Type	74.5%
dc.Language	74.6%
dc.Date	100.0%
dc.Format	100.0%
dc.Identifier	100.0%
dc.Title	100.0%

Element view



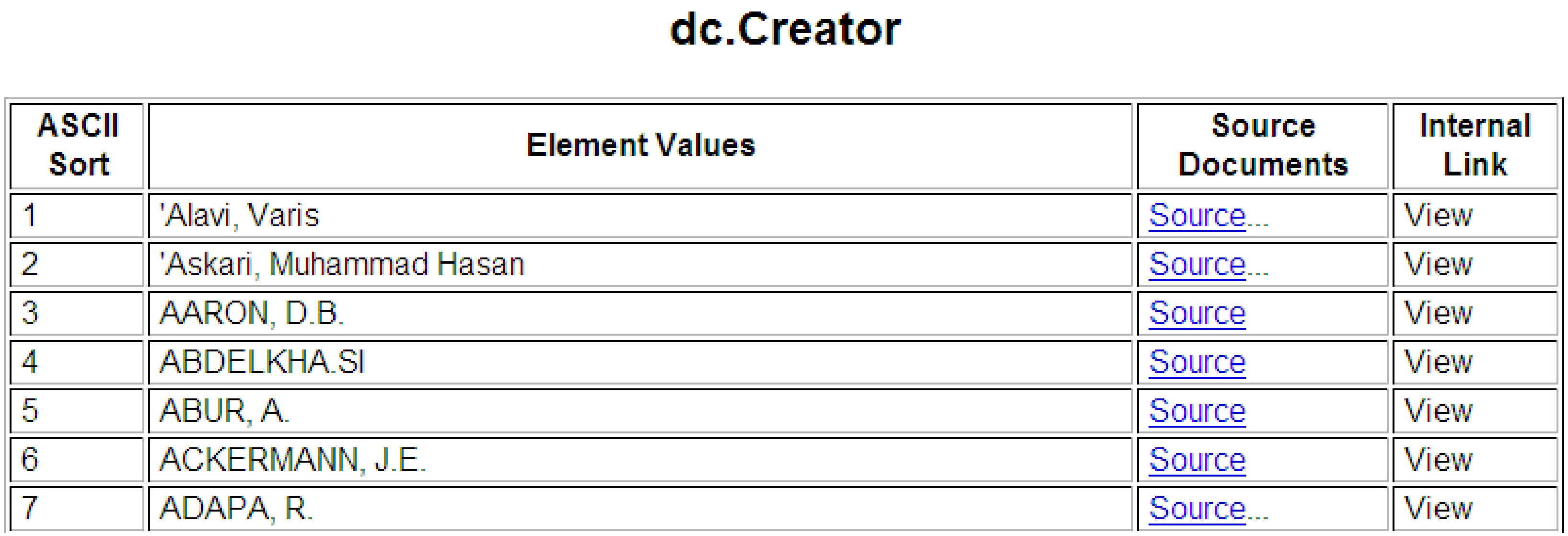
Metadata Element Detail:dc.Title

Total Number of Records	6015
Unique Values	5794
Total times element used	6028
No. of records containing element	6015
Completeness	100.0%
Minimum dc.Title usage in any record What's this?	1
Maximum dc.Title usage in any record What's this?	2
Average dc.Title usage/record What's this?	1.0
Mode of dc.Title usage/record What's this?	1
Coverage of the mode of dc.Title usage/record What's this?	99.8%
View Potential Duplicate List	No Records Missing dc.Title
View Full Frequency Sorted list	View Full ASCII Sorted list

ASCII-Based	First Five
1	"Allah Hafiz"
2	"As Bad as All That!"
3	"Deconstructing" a "Deconstructionist" Urdu Story: "Ek Kahan ...
4	"Hic Facet Arthurus, Rex Quondam, Rexque Futurus:" The Analy ...
5	"Hit It With a Stick and It Won't Die": Urdu Language, Musli ...
......	Last Five
5790	to W. A. Sredenschek in praise of recent address to New York ...
5791	to W. A. Sredenschek re: L. D. Miles' and Dick Bradshaw's pr ...
5792	to W. A. Sredenschek re: success of meeting with Control Div ...
5793	The Ghat of the Only World: Agha Shahid Ali in Brooklyn
5794	'Seeing' song in Bollywood : landscape, the postnational, an ...

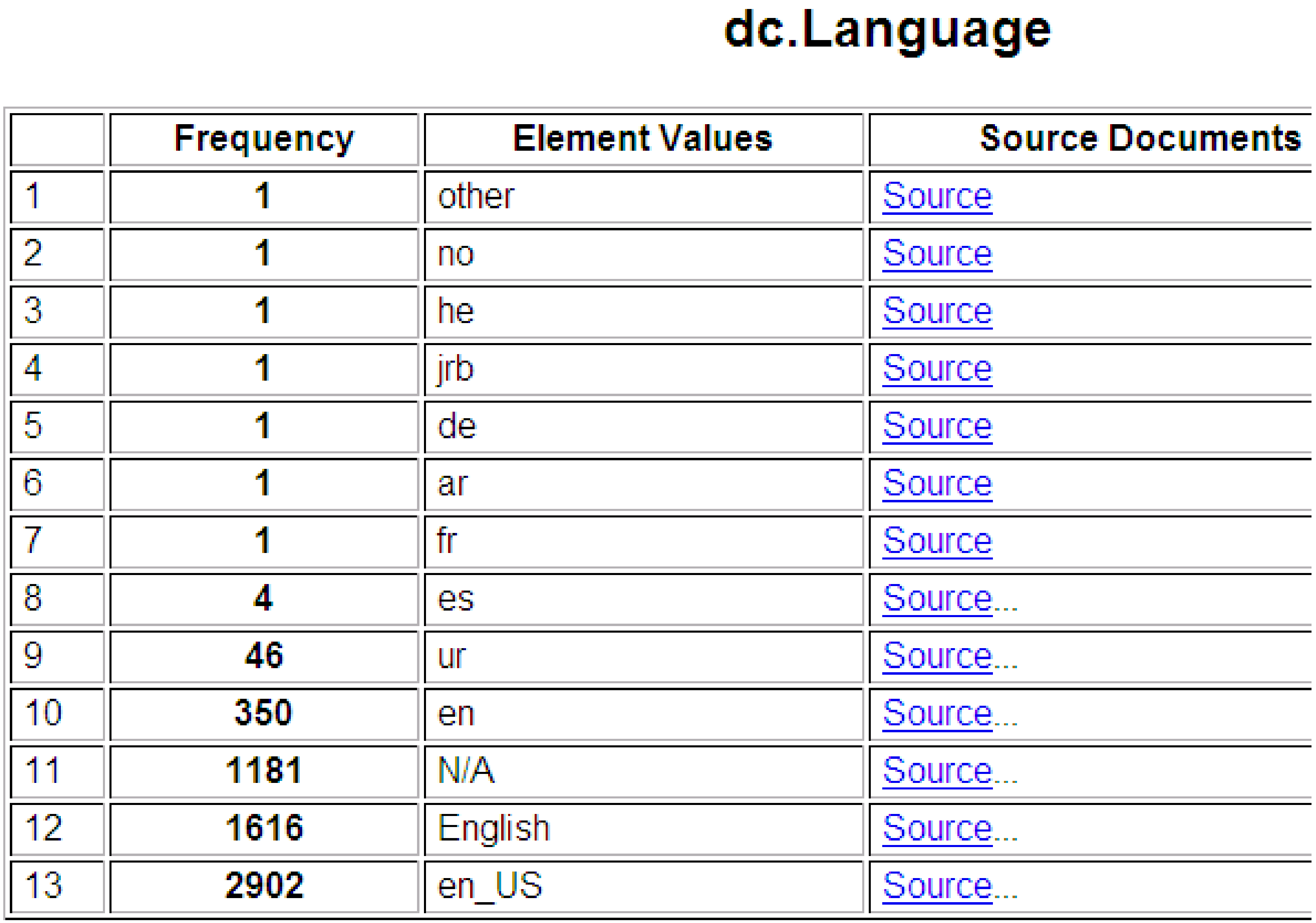
Frequency-Based:	First Five
1. (No. of occurrences: 1)	Feminist Collections, v.12, no.1 (fall 1990)
2. (No. of occurrences: 1)	A root of less evil

ASCII sorted element list



dc.Creator

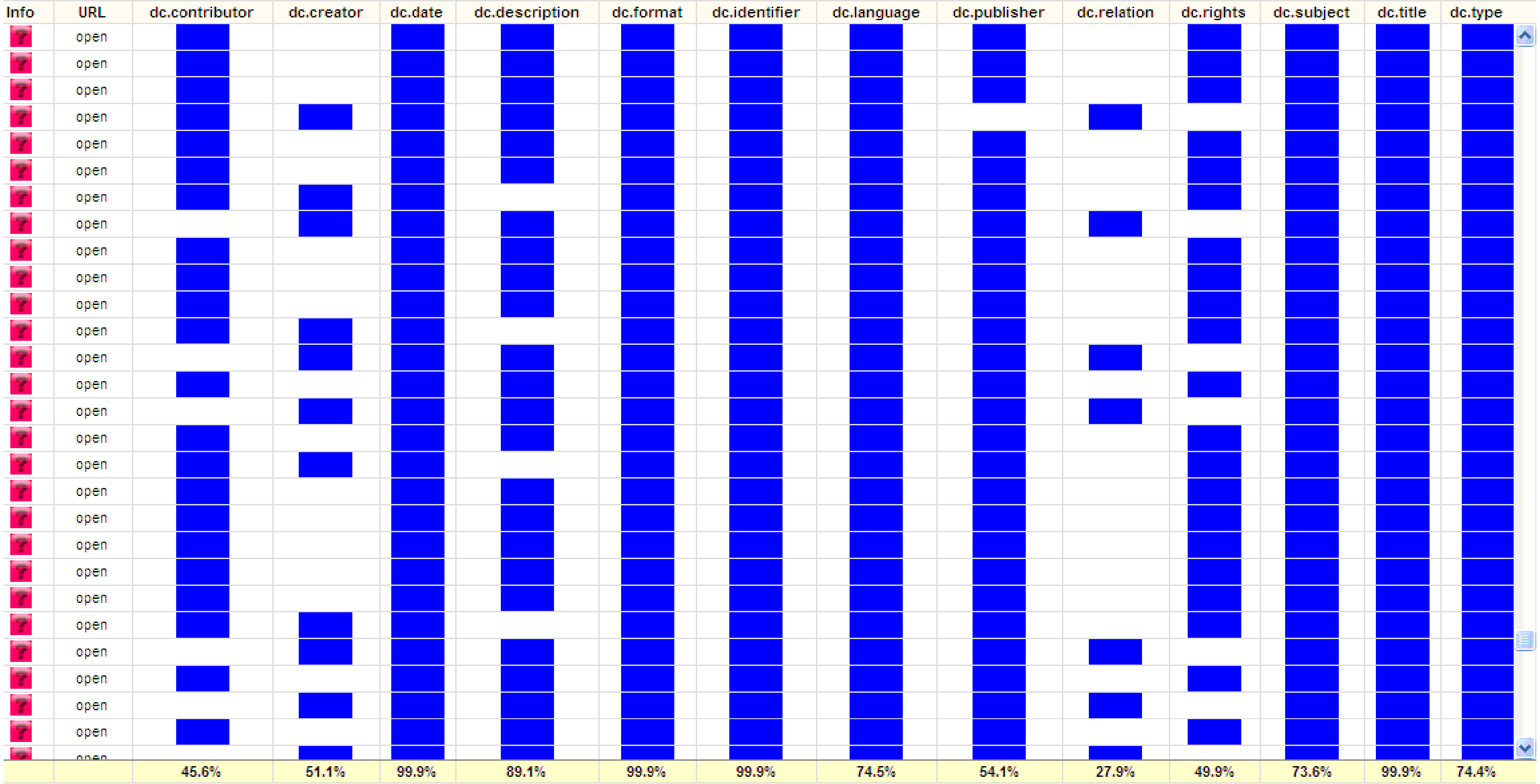
ASCII Sort	Element Values	Source Documents	Internal Link
1	'Alavi, Varis	Source...	View
2	'Askari, Muhammad Hasan	Source...	View
3	AARON, D.B.	Source	View
4	ABDELKHA.SI	Source	View
5	ABUR, A.	Source	View
6	ACKERMANN, J.E.	Source	View
7	ADAPA, R.	Source...	View



dc.Language

	Frequency	Element Values	Source Documents
1	**1**	other	Source
2	**1**	no	Source
3	**1**	he	Source
4	**1**	jrb	Source
5	**1**	de	Source
6	**1**	ar	Source
7	**1**	fr	Source
8	**4**	es	Source...
9	**46**	ur	Source...
10	**350**	en	Source...
11	**1181**	N/A	Source...
12	**1616**	English	Source...
13	**2902**	en_US	Source...

Visualisation



Info	URL	dc.contributor	dc.creator	dc.date	dc.description	dc.format	dc.identifier	dc.language	dc.publisher	dc.relation	dc.rights	dc.subject	dc.title	dc.type
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■	■	■	■	■	■	■		■		■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■	■	■		■	■	■	■		■	■	■	■
?	open		■	■	■	■	■	■	■	■		■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■	■	■		■	■	■	■		■	■	■	■
?	open		■	■	■	■	■	■	■	■		■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open		■	■	■	■	■	■	■	■		■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■	■	■		■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open	■	■	■		■	■	■	■		■	■	■	■
?	open		■	■	■	■	■	■	■	■		■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open		■	■	■	■	■	■	■	■		■	■	■
?	open	■		■	■	■	■	■	■		■	■	■	■
?	open		■	■	■	■	■	■	■	■		■	■	■
		45.6%	51.1%	99.9%	89.1%	99.9%	99.9%	74.5%	54.1%	27.9%	49.9%	73.6%	99.9%	74.4%

MINDS ≈ 6000 records → 6,000 row HTML table → browser stress

Qualitative Feedback

- Online Survey, interviews, blog comments, email…
- + ve:
  - “really useful”, “I found and used this tool last week and found it very useful for exploring our own repository”
  - “it is so nice to see people working in this area.”
  - Observed interactive error detection & correction
- - ve
  - Not enough linking back to repositories
  - slow, not stable
- other:
  - Individual differences in preference for the text/table output v visualisation

Issues

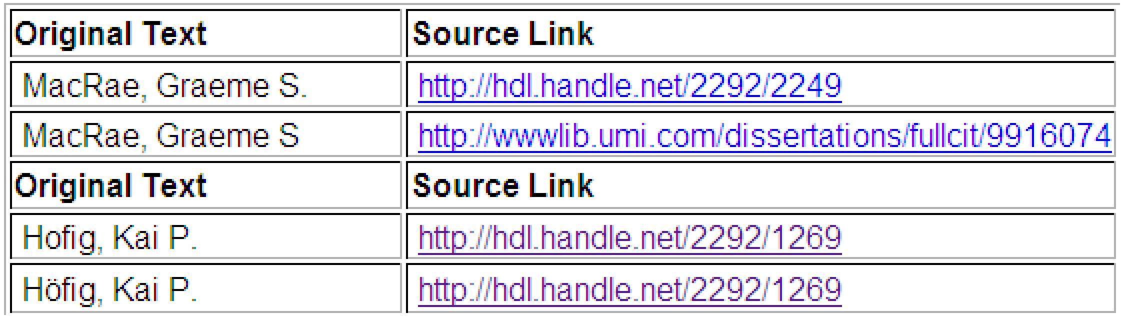
- Discovered GS3 bugs that only occur with large numbers (> 150) of collections
- No incremental harvesting yet
  - So download *all* of the repository metadata every time = slow
- Most IRs don’t expose their *qualified* Dublin Core
  - … so we can’t analyse it
  - The resulting ‘dumbing down’ mixes datatypes (text, URLs, DOIs etc.) and lowers the effectiveness of analysis
- Some expose other types of metadata:
  - METS, MODS, ETDs etc

Beta version: features

- lists of potential duplicate values for each element
  - using approximate string matching (edit distances)
- lists of records that are missing particular elements
- better linking to source item records
- greatly improved stability

Potential Duplicates

- Based on Levenshtein edit distance
  - added custom variable costs (quotes)
  - & custom tweaks (e.g. case folding) & threshold = 2



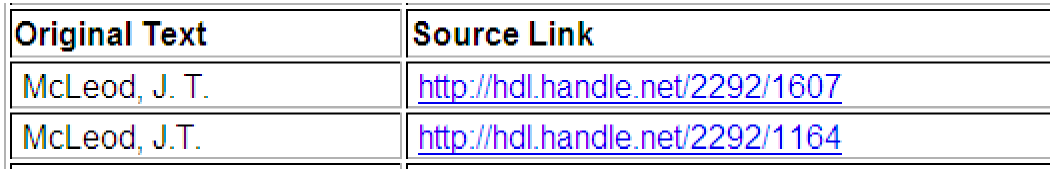
punctuation

Original Text	Source Link
MacRae, Graeme S.	http://hdl.handle.net/2292/2249
MacRae, Graeme S	http://wwwlib.umi.com/dissertations/fullcit/9916074

diacritic

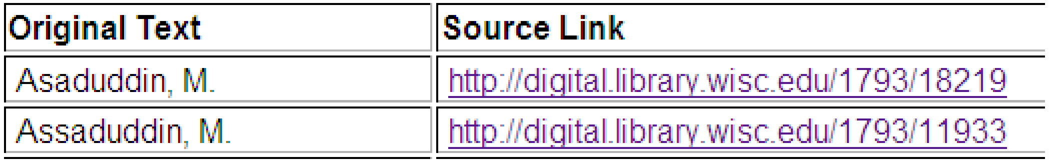
Original Text	Source Link
Hofig, Kai P.	http://hdl.handle.net/2292/1269
Höfig, Kai P.	http://hdl.handle.net/2292/1269

spacing



Original Text	Source Link
McLeod, J. T.	http://hdl.handle.net/2292/1607
McLeod, J.T.	http://hdl.handle.net/2292/1164

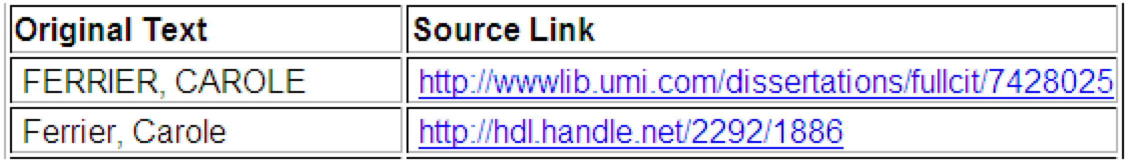
typos



Original Text	Source Link
Asaduddin, M.	http://digital.library.wisc.edu/1793/18219
Assaduddin, M.	http://digital.library.wisc.edu/1793/11933

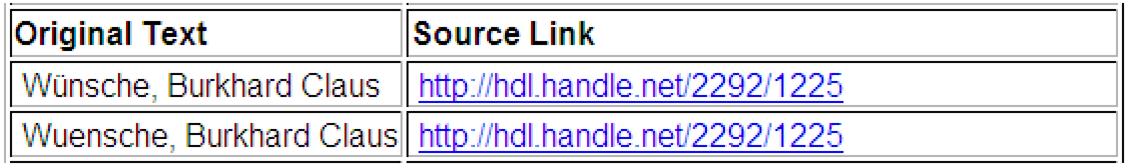
Duplicate detection examples

capitalisation



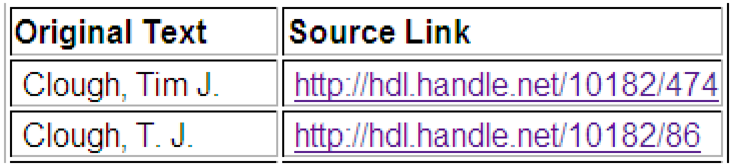
Original Text	Source Link
FERRIER, CAROLE	http://wwwlib.umi.com/dissertations/fullcit/7428025
Ferrier, Carole	http://hdl.handle.net/2292/1886

diacritic transliteration



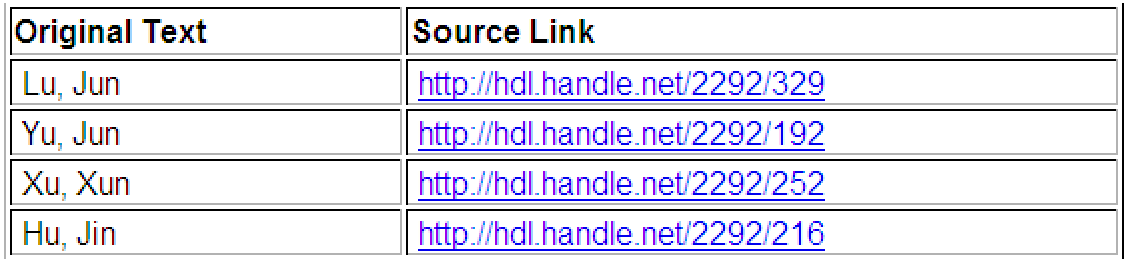
Original Text	Source Link
Wünsche, Burkhard Claus	http://hdl.handle.net/2292/1225
Wuensche, Burkhard Claus	http://hdl.handle.net/2292/1225

initialisation



Original Text	Source Link
Clough, Tim J.	http://hdl.handle.net/10182/474
Clough, T. J.	http://hdl.handle.net/10182/86

non-duplicates

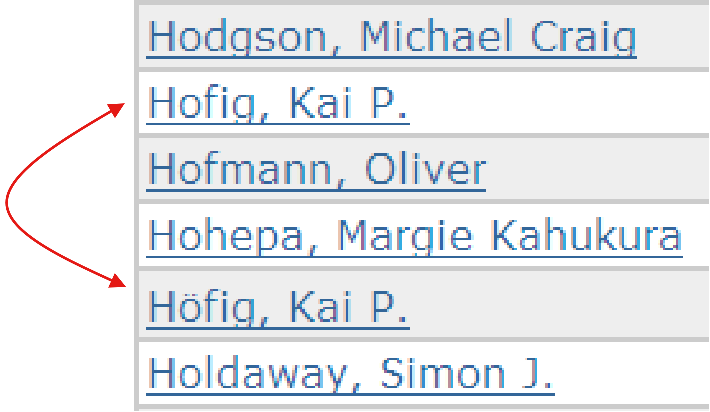


Original Text	Source Link
Lu, Jun	http://hdl.handle.net/2292/329
Yu, Jun	http://hdl.handle.net/2292/192
Xu, Xun	http://hdl.handle.net/2292/252
Hu, Jin	http://hdl.handle.net/2292/216

Duplicate detection: initial findings

- Lots of small differences
  - Spacing, punctuation, accents, quotes etc
- Not many ‘big’ errors
  - But only a small sample and we lack local knowledge of collections – but errors are there
- Some types of data don’t produce useful results
  - URLs, DOIs, dates, filesizes, sequences (Part 1, Part 2.., years)
- Consequences of differences depend on the (IR) software

Examples from the IR browsing structures



Hodgson, Michael Craig

Hofig, Kai P.

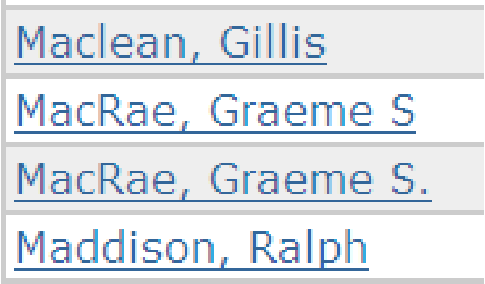
Hofmann, Oliver

Hohepa, Margie Kahukura

Höfig, Kai P.

Holdaway, Simon J.

"encoded character" problem

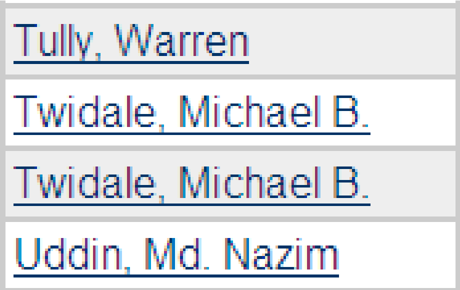


Maclean, Gillis

MacRae, Graeme S

MacRae, Graeme S.

Maddison, Ralph

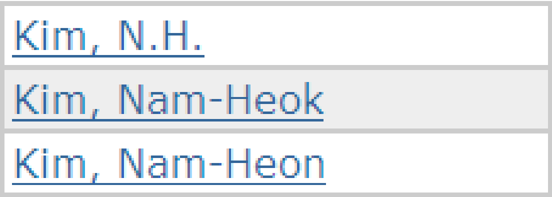


Tully, Warren

Twidale, Michael B.

Twidale, Michael B.

Uddin, Md. Nazim



Kim, N.H.

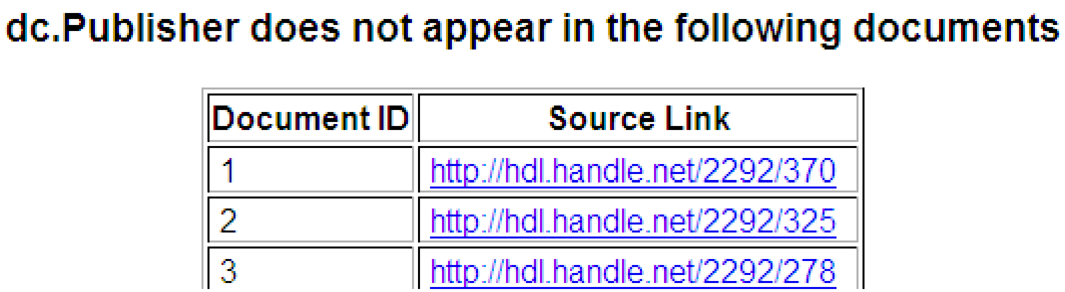
Kim, Nam-Heok

Kim, Nam-Heon

"Authority control is desperately needed for metadata"

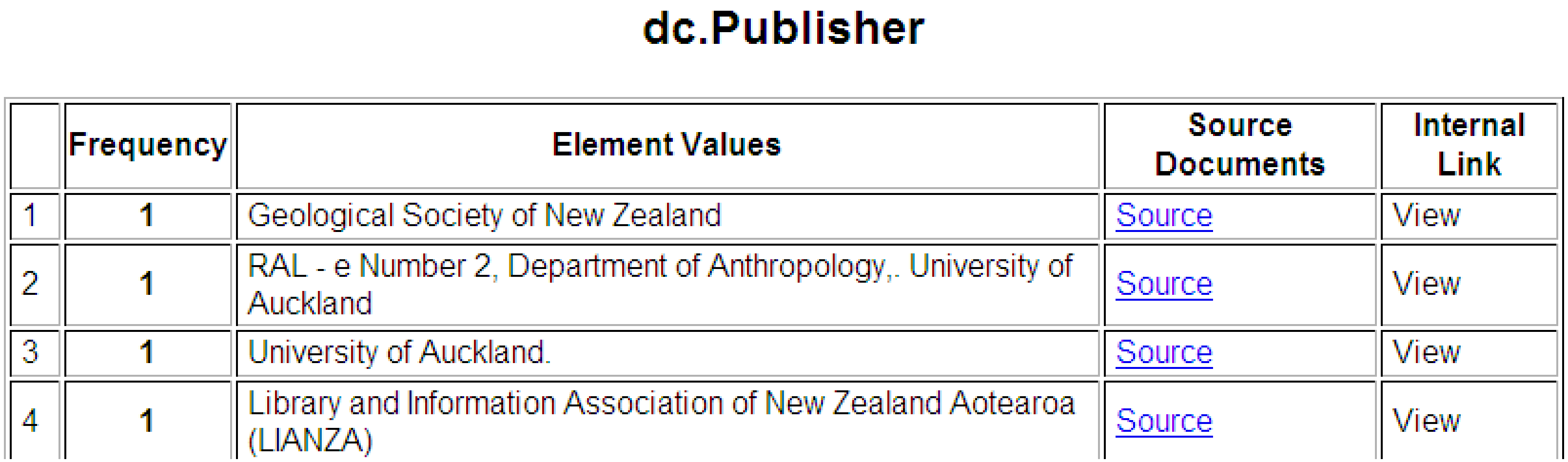
http://wiki.dspace.org/index.php/LessonsLearned

Other improvements



dc.Publisher does not appear in the following documents

Document ID	Source Link
1	http://hdl.handle.net/2292/370
2	http://hdl.handle.net/2292/325
3	http://hdl.handle.net/2292/278



dc.Publisher

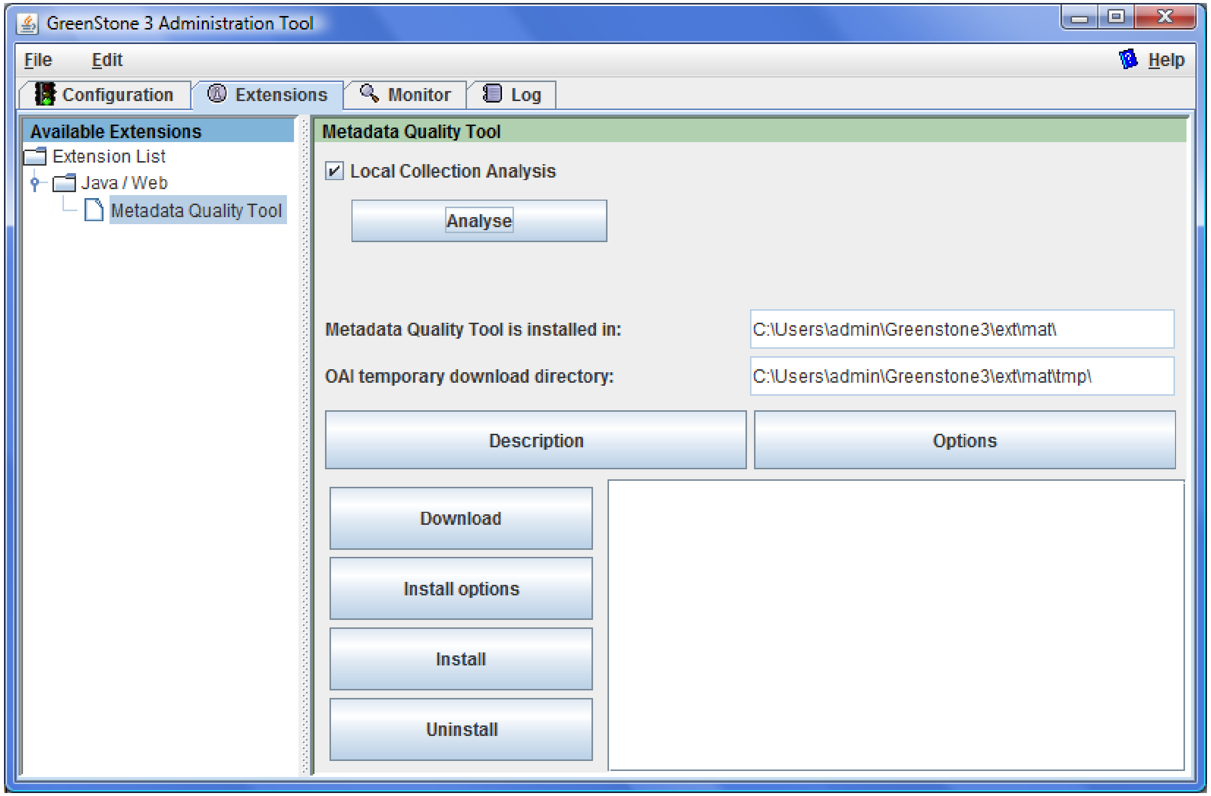
	Frequency	Element Values	Source Documents	Internal Link
1	1	Geological Society of New Zealand	Source	View
2	1	RAL - e Number 2, Department of Anthropology,. University of Auckland	Source	View
3	1	University of Auckland.	Source	View
4	1	Library and Information Association of New Zealand Aotearoa (LIANZA)	Source	View

Next Steps

- Running it as a public service is a lot of work
  - Especially as code base changes daily
  - Remote data sources aren't always valid
  - Dealing with our security conscious Technical Support people is interesting
- Not very efficient to have the world using one service
  - No incremental harvesting
  - No incremental collection building
- Reports are public (guess URLs)
  - Not ideal for systems in development
- So...

Add to every GS3 installation

- Add an extensions mechanism to GS3
- Mat is the first working extension

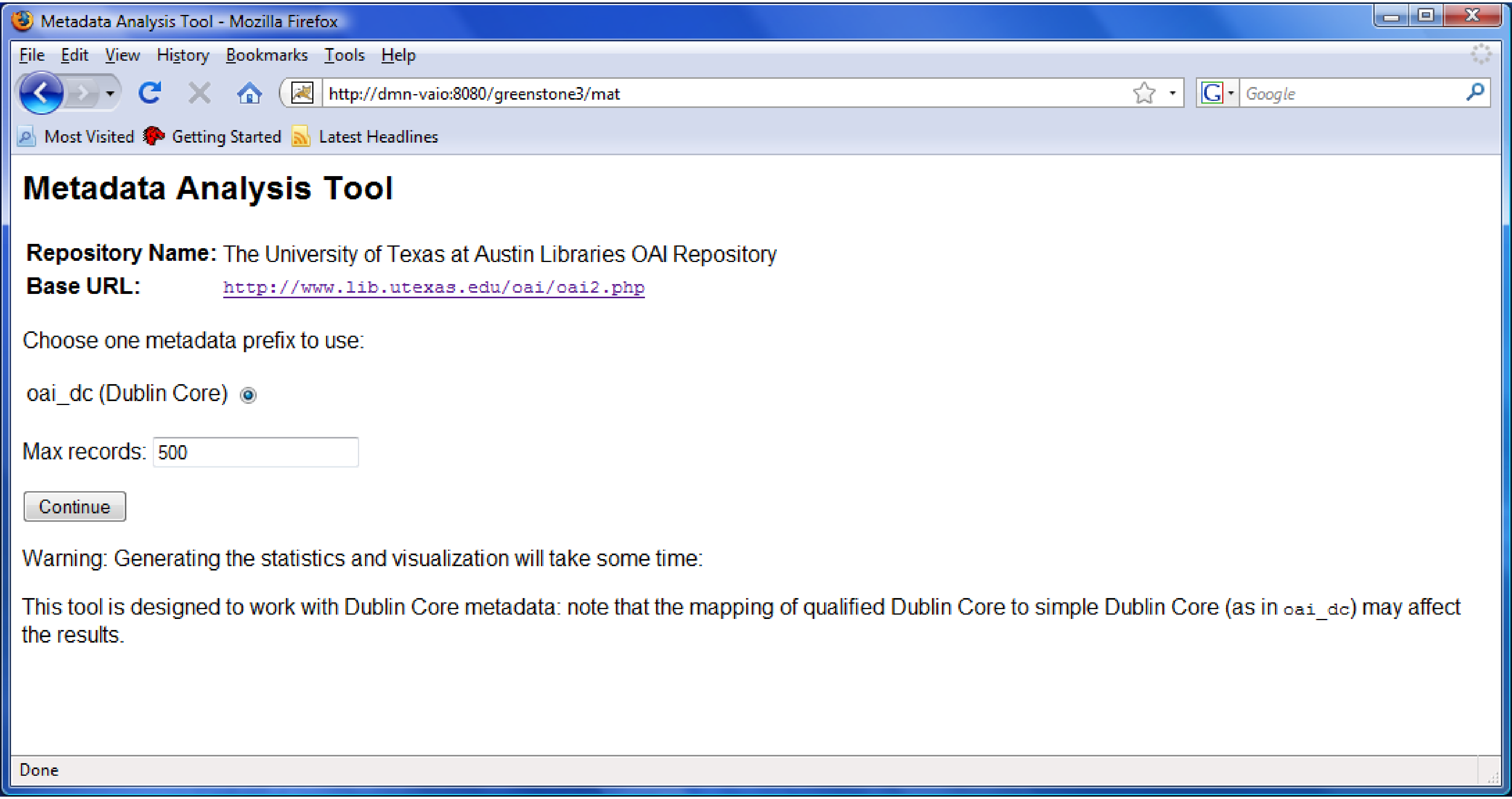


Metadata Quality Tool

☑ Local Collection Analysis

Metadata Quality Tool is installed in: C:\Users\admin\Greenstone3\ext\mat\

OAI temporary download directory: C:\Users\admin\Greenstone3\ext\mat\tmp\


Metadata Analysis Tool - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://dmn-vaio:8080/greenstone3/mat
Google
Most Visited
Getting Started
Latest Headlines

Metadata Analysis Tool

Repository Name: The University of Texas at Austin Libraries OAI Repository

Base URL: http://www.lib.utexas.edu/oai/oai2.php

Choose one metadata prefix to use:

oai_dc (Dublin Core) ●

Max records: 500

Warning: Generating the statistics and visualization will take some time:

This tool is designed to work with Dublin Core metadata: note that the mapping of qualified Dublin Core to simple Dublin Core (as in `oai_dc`) may affect the results.

Aim

- Every new Greenstone install will have a metadata analysis tool built-in
- Every new Greenstone install will have a simple method to deploy a Mat tool
  - Either just for themselves, or
  - for limited (e.g. by domain) or public use

KRIS – Kiwi Research Information Service

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

- nzresearch.org.nz

Project Goal

To build a national discovery service for the research held in institutional repositories in New Zealand, for the mutual benefit of researchers, research users, and research institutions.

Some KRIS material by Gordon Paynter

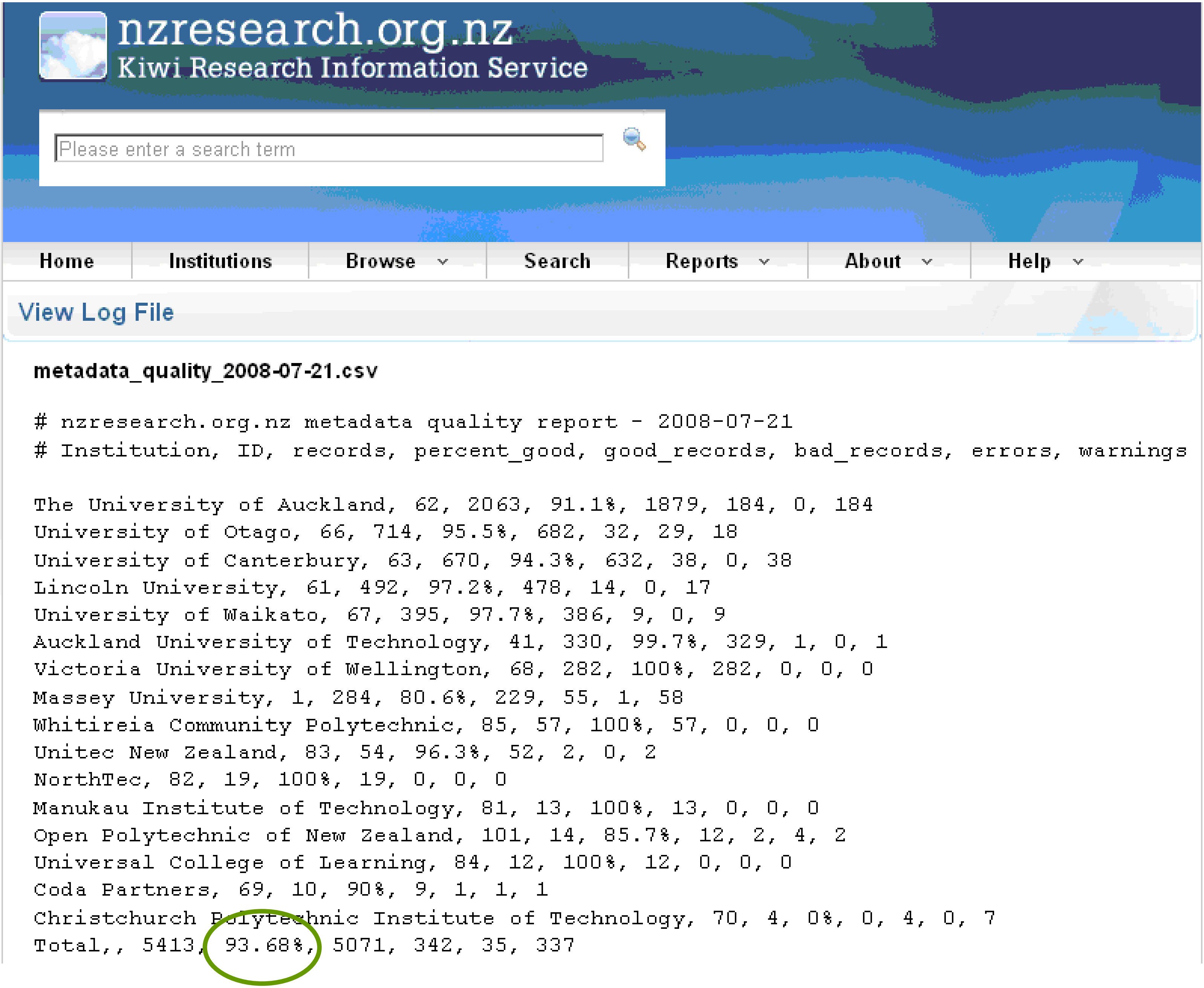
Characteristics

- All universities (8) in the country
  - Some of the Technology Institutes/Polytechnics/Wananga (23)
- Agreed metadata policies
  - Based on unqualified DC, low entry level (only 4 mandatory elements)
  - http://www.natlib.govt.nz/catalogues/library-documents/national-research-discovery-service-metadata-guidelines
  - Validate harvested metadata against policies
- Custom software
  - Oracle DB backend, XSL for validation

Input & Output

- Overnight incremental harvests
- Disseminates via CSV, OAI, RSS
  - Download OAI error set at any time
  - RSS by author, subject
  - RSS metadata errors
    - IR administrator can receive daily RSS error feed about their own repository
- “Repositories don't have to do anything... it will just work” Paynter (2007)

Demo



View Log File

metadata_quality_2008-07-21.csv

```
# nzresearch.org.nz metadata quality report - 2008-07-21
# Institution, ID, records, percent_good, good_records, bad_records, errors, warnings

The University of Auckland, 62, 2063, 91.1%, 1879, 184, 0, 184
University of Otago, 66, 714, 95.5%, 682, 32, 29, 18
University of Canterbury, 63, 670, 94.3%, 632, 38, 0, 38
Lincoln University, 61, 492, 97.2%, 478, 14, 0, 17
University of Waikato, 67, 395, 97.7%, 386, 9, 0, 9
Auckland University of Technology, 41, 330, 99.7%, 329, 1, 0, 1
Victoria University of Wellington, 68, 282, 100%, 282, 0, 0, 0
Massey University, 1, 284, 80.6%, 229, 55, 1, 58
Whitireia Community Polytechnic, 85, 57, 100%, 57, 0, 0, 0
Unitec New Zealand, 83, 54, 96.3%, 52, 2, 0, 2
NorthTec, 82, 19, 100%, 19, 0, 0, 0
Manukau Institute of Technology, 81, 13, 100%, 13, 0, 0, 0
Open Polytechnic of New Zealand, 101, 14, 85.7%, 12, 2, 4, 2
Universal College of Learning, 84, 12, 100%, 12, 0, 0, 0
Coda Partners, 69, 10, 90%, 9, 1, 1, 1
Christchurch Polytechnic Institute of Technology, 70, 4, 0%, 0, 4, 0, 7
Total,, 5413, 93.68%, 5071, 342, 35, 337
```

“State of the nation’s metadata” (Paynter 2007)

Examples of errors

Error	Record has no Date	4
Error	Record has no HTTP URL	3
Error	Record has no Author	3
Error	Record has no Title	1
Warning	Unknown Type value: NonPeerReviewed	629
Warning	Unknown Type value: journal	128
Warning	Unknown Type value: PeerReviewed	40
Warning	Author not in “Citename, Firstnames” format	8
Warning	Unknown Type value: Book Section	3

Paynter (2007)

Experiences of running services

- Controlled population of KRIS is much easier than public demand-driven Mat
- Some IR administrators know less than you might expect:
  - "Enter the OAI URL of your repository"
- Also, may not have technical control over what is harvested
  - Even if they wanted to, they can't turn on qualified DC export via OAI-PMH
- Even division between preference for textual outputs v visualisations
- Web-based tools work well with web-based IR admin interfaces

Experiences 2

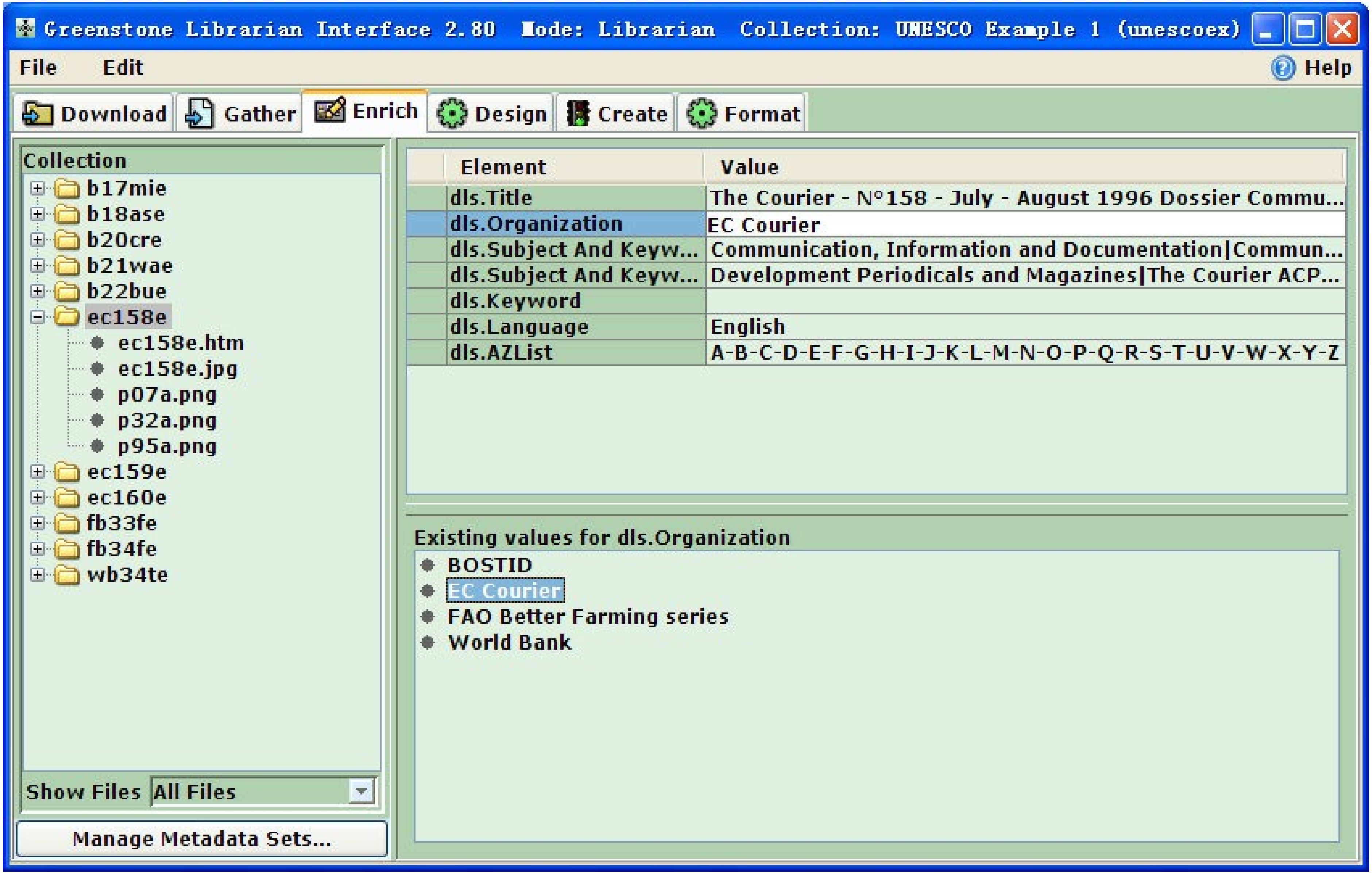
- Overnight zero-effort updates are appreciated
  - Works because of a known small population of repositories (KRIS)
  - To work for Mat we would have to harvest everything or turn to an account-based system
- KRIS was purpose-built, whereas Mat attempts to leverage existing GS technologies
  - And it shows
- Security issues in being demand-led, KRIS's fixed population is much safer

Summary

- Metadata assessment tools are needed
- Existing collection software doesn’t help much
    - No authority control
    - Poor feedback
    - No built-in analysis tools
- Much more to do
    - Type-aware – parsing data formats (Dates, URLs, IMT etc) (Mat)
    - Efficiency and stability (Mat)
    - More agreed standards (KRIS)

Requirements for metadata processing

- Built-in set of types
  - IMT, DCMI-types, language codes
- Built-in set of patterns
  - URLs, DOIs, ISBNs, IMT extensions
- User specified patterns
  - REs, Java methods, JS functions
- Use these types & patterns for
  - Input into analysis tool, Metadata Entry in Enrich Panel
- Metadata Cleaning/Processing
  - Java String API – without needing to write any Java
  - REs, custom JS functions
  - (done but not yet integrated)



Show Files: All Files

Element	Value
dls.Title	The Courier - N°158 - July - August 1996 Dossier Commu...
dls.Organization	EC Courier
dls.Subject And Keyw...	Communication, Information and Documentation|Commun...
dls.Subject And Keyw...	Development Periodicals and Magazines|The Courier ACP...
dls.Keyword	
dls.Language	English
dls.AZList	A-B-C-D-E-F-G-H-I-J-K-L-M-N-O-P-Q-R-S-T-U-V-W-X-Y-Z

Existing values for dls.Organization

- BOSTID
- EC Courier
- FAO Better Farming series
- World Bank

Questions?

nzdl.org/greenstone3/mat
nzresearch.org.nz

Nichols, D.M., Chan, C-H., Bainbridge, D., McKay, D. and Twidale, M.B. (2008) A lightweight metadata quality tool, *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'08)*. 385-388.

Nichols, D.M., Paynter, G.W., Chan, C-H., Bainbridge, D., McKay, D., Twidale, M.B. and Blandford, A. (2009) Experiences in deploying metadata analysis tools for institutional repositories *Cataloging and Classification Quarterly* 47(3/4) 229-248.

cs.waikato.ac.nz/~daven