THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Metadata Qualities for Digital Libraries

Chu-Hsiang Chan

This thesis is submitted in partial fulfillment of the requirements for the degree of Master of Science at the University of Waikato.

August 2008

Abstract

The quality of metadata in a digital library is an important factor in ensuring access for end-users. Several studies have tried to define quality frameworks and assess metadata but there is little user feedback about these in the literature. As collections grow in size maintaining quality through manual methods becomes increasingly difficult for repository managers.

This research presents the design and implementation of a web-based metadata analysis tool for digital repositories. The tool is built as an extension to the Greenstone3 digital library software.

We present examples of the tool in use on real-world data and provide feedback from repository managers. The evidence from our studies shows that automated quality analysis tools are useful and valued service for digital libraries.

Acknowledgements

This project has been undertaken as part of Greenstone Digital Library Research Group and I would like to thank everyone who helped me in this research project.

I want to acknowledge the specific contributions of:
David Bainbridge: For improvement to Greenstone OAI handling.
Dana McKay: For interviews of repository managers
Katherine Don and the Greenstone programmers: For bug fixes to the Greenstone3 code.

I also want to thank everyone who completed our survey or interview to help us improve the tool. Especially, I would like to give my thanks to my supervisor Dr. Dave Nichols. Dr. Dave Nichols always supported me in my research work and helped me complete this work.

Table of Contents

Chapter 1. Introduction

Chapter 2. Background

Chapter 3. System Design and Implementation

Chapter 4. Mat Evolution

Chapter 5. Discussion

Chapter 6. Conclusion and Future Work

Appendices

List of Figures

List of Tables

Chapter 1

Introduction

As digital collections grow in size, the usability and effectiveness of digital libraries have become important issues for repository managers. The usability of digital repositories for end-users is simply affected by the quality of metadata records. Therefore, digital repository managers need to ensure that poor metadata quality does not affect user satisfaction. For a small-size repository, the metadata quality can be controlled by reviewing metadata records manually. However, this kind of human review approach is infeasible for large collections. Therefore, more and more repository managers are looking for useful metadata quality tools to ensure the quality of their collections. The aim of this research is to explore the requirements for these tools by building a practical analysis service.

1.1 Motivation

It is difficult for repository managers to manually maintain quality for large collections. If repository managers cannot keep metadata quality high, resources in digital collections may not be accessible to end-users. Typically, this poor metadata quality problem can be divided into six categories: (Stvilia et al., 2004)

- lack of completeness
- redundant metadata
- lack of clarity

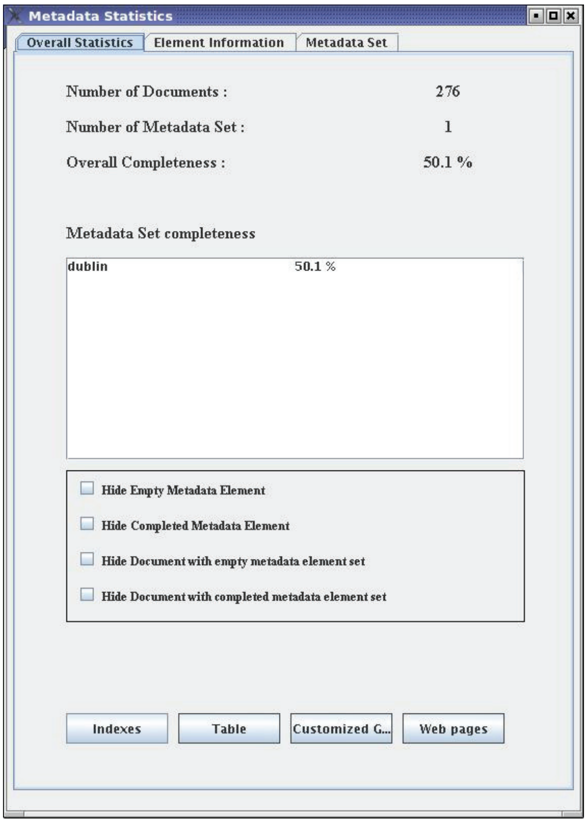
- incorrect use of metadata schema or semantic inconsistency
- structural inconsistency
- inaccurate representation

Most of these categories can be addressed by both computer software and human experts. However, human review approaches may not be consistent in some cases. Using metadata quality tools we can make ensure that every record is treated equally. Most digital library software does not have effective quality assurance mechanisms for their metadata. Digital repository managers often need to check each metadata record manually in order to find quality problems. To address this problem, we have developed a software application to help repository managers understand their digital collections.

1.2 Description

The original goal of this research project was to design a tool that could analyse Greenstone (Bainbridge et al., 2004) collections and generate statistical reports and visualisations for end-users to revise their metadata. The statistical report provides end-users with detailed information of the usage of each metadata element, the numbers of each element used, and a list of metadata element values. The visualisation tool provides an overview of data sets and shows distribution of data points.

Several metadata quality tools have been developed, but there is little user feedback about these in the literature. In order to get feedback from both end-users and repository managers about this metadata quality tool, we developed a new online metadata quality application to achieve this goal. The core of the main program is written in Java, chosen because the Greenstone digital library is also written in Java. This web-based tool is constructed on top of Greenstone. Therefore, repository managers can use this metadata quality tool without installing any computer software. Figure 1 is a screenshot of a metadata statistical summary of a collection from the Java version. The online tool is known as Mat for Metadata Analysis Tool.



Number of Documents : 276

Number of Metadata Set : 1

Overall Completeness : 50.1 %

Metadata Set completeness

dublin 50.1 %

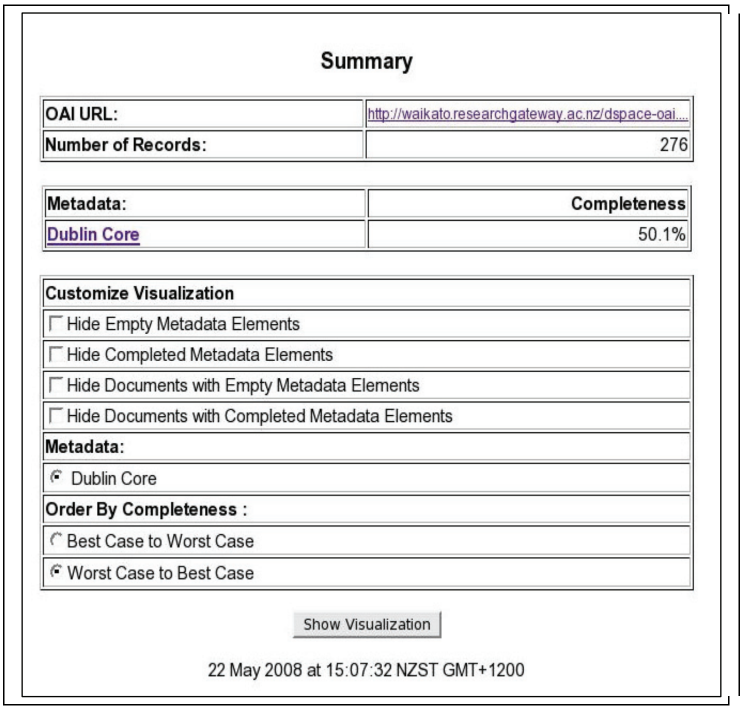
□ Hide Empty Metadata Element

□ Hide Completed Metadata Element

□ Hide Document with empty metadata element set

□ Hide Document with completed metadata element set

Figure 1 : Overall metadata statistics of a collection (Java version)



Summary

OAI URL:	http://waikato.researchgateway.ac.nz/dspace-oai...
Number of Records:	276

Metadata:	Completeness
Dublin Core	50.1%

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

22 May 2008 at 15:07:32 NZST GMT+1200

Figure 2 : A screen shot of the summary page of Mat (web version)

The online version (see Figure 2) requires repository managers to specify the URL of a repository and wait for Greenstone to download metadata records and build collections. The main advantage of the online version is that repository managers are not required to install any software but it takes longer to complete the tasks. The Java version provides more functionality and allows end-users to create their own visualisations.

The main contributions of this thesis are:

- Building a metadata analysis tool that integrates with Greenstone.
- Creating a web version of the tool (Nichols and Chan et al., 2008).
- Exploring requirements for metadata tools through feedback from repository managers (Nichols and Paynter et al., 2008).

This project has been undertaken as part of Greenstone Digital Library Research Group and we wish to acknowledge the specific contributions of:

- David Bainbridge: For improvement to Greenstone OAI handling.
- Dana McKay: For interviews of repository managers
- Katherine Don and the Greenstone programmers: For bug fixes to the Greenstone3 code.

1.3 Thesis Structure

The remaining parts of this thesis are organised as follows: Chapter 2 gives the background on digital libraries and discusses related literature. Chapter 3 describes the software design, development and implementation issues. Chapter 4 presents the evolution of the Mat tool. In Chapter 5, the statistical analysis is described and qualitative feedback is presented. Finally, the conclusion and future research are given.

Chapter 2

Background

2.1 Introduction

In this chapter, we discuss the technology of digital libraries including software systems for building collections, standards, and protocols. We describe studies on the quality of metadata in digital libraries and their implications for metadata tools.

2.2 Digital Libraries

As a result of the development of information and communication technology, more and more digital resources such as text, document, video and music have been produced. Generally speaking, they are usually stored in the computers and spread by the Internet. The term digital library was first popularized by the NSF/DARPA/NASA Digital Libraries Initiative in 1994 (Fox, 1999).

> *"Digital libraries are organized collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible" (Lesk, 1997).*

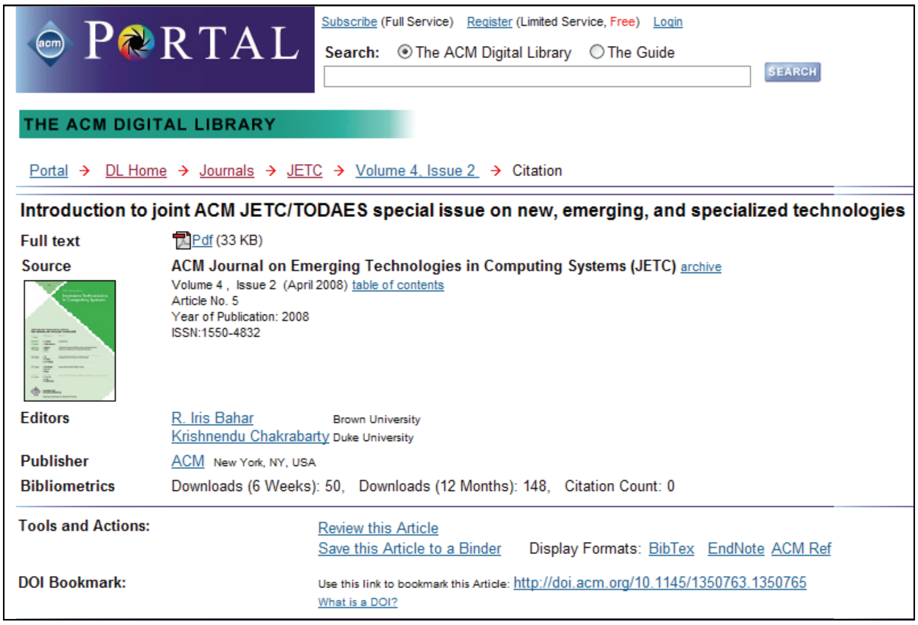
Digital Library is a type of information retrieval system and mainly used in preserving digital resources. To be considered as a digital library, a collection of information must be accessible to both humans and computer systems. The most obvious advantage of digital libraries is that people can more easily and rapidly access resources. Three main advantages of digital library are listed below:

- Space: A traditional library requires lots of storage space while digital storage can save a considerable amount of space.

- No physical boundary: Users would not have to go to the library physically and they can use computers to access the information.

- Informational retrieval: Users can use key phrases, titles, and names to search the collection and receive both metadata and content immediately. There is a software application similar to digital libraries named OPAC (Online Public Access Catalog). OPAC allows end-users to search book titles, keywords and authors but they have to go the library physically to view the document content. Digital libraries allow end-users to view the documents on the Internet.

There are several well-known digital libraries such as Project Perseus (http://www.perseus.tufts.edu/) and Project Gutenberg (http://www.gutenberg.org/wiki/Main_Page) around the world. In addition, many academic institutions are currently building institutional repositories for content such as books, papers, theses and etc. Many academic repositories are available to the public with some restrictions. A personal digital library could be used to store pictures, music, electronic books and films. Digital libraries typically provide search and browse interfaces which allow users to locate digital materials. Figure 3 shows the searching interface of the ACM (http://portal.acm.org/portal.cfm) digital library showing relevant metadata. Since it is difficult for human to build digital collections, we consider computer software that could do this job for us. In the following sections, three software systems for building digital libraries will be introduced.



Subscribe (Full Service) Register (Limited Service, Free) Login

Search: ● The ACM Digital Library ○ The Guide

THE ACM DIGITAL LIBRARY

Portal → DL Home → Journals → JETC → Volume 4, Issue 2 → Citation

Introduction to joint ACM JETC/TODAES special issue on new, emerging, and specialized technologies

Full text	Pdf (33 KB)
Source	ACM Journal on Emerging Technologies in Computing Systems (JETC) archive Volume 4 , Issue 2 (April 2008) table of contents Article No. 5 Year of Publication: 2008 ISSN:1550-4832
Editors	R. Iris Bahar Brown University Krishnendu Chakrabarty Duke University
Publisher	ACM New York, NY, USA
Bibliometrics	Downloads (6 Weeks): 50, Downloads (12 Months): 148, Citation Count: 0

Tools and Actions: Review this Article
Save this Article to a Binder Display Formats: BibTex EndNote ACM Ref

DOI Bookmark: Use this link to bookmark this Article: http://doi.acm.org/10.1145/1350763.1350765
What is a DOI?

Figure 3 : ACM digital library

2.2.1 Greenstone

Greenstone (http://www.greenstone.org/) is a software system for building digital collections under the GNU General Public License (Witten and Bainbridge, 2003). It is not a digital library but a tool for building digital libraries. It assists end-users to organise information and distribute it via the web. Greenstone is produced by the New Zealand Digital Library Project at the University of Waikato.

The Greenstone Librarian Interface (GLI) is a Java-based graphical user interface for repository administrators to manage their collections (See Figure 4). Repository managers can use it to import documents, edit metadata values, build new collections, and distribute collections to end-users. GLI has four modes: Librarian Assistant, Librarian, Library Systems Specialist, and Expert. Modes control the level of detail within the interface.

Figure 5 is a screenshot of main interface for end-users to find documents. There are two main approaches to retrieve information in Greenstone collections: Searching and Browsing.

- Searching: Greenstone constructs full-text indexes from the document text. Hence, the system allows end-users to search for particular words that appear in the text of a document. This facility is particularly useful for approximate search.

- Browsing: End-users can browse documents by titles or subjects. The system lists all available authors and subjects for end-users to find the information easily. Figure 5 shows four approaches for end-users to locate a resource. They can search for a key word or browse the “titles”, “subjects”, and “organisations” to find the objects. These classifiers are changeable and depended on users needs.

Greenstone3 (Buchanan et al., 2005) is a new implementation designed to improve the flexibility, modularity and extensibility of the original Greenstone digital library software. The original Greenstone project (Greenstone2) is a complete digital library that facilitates creation, organisation and management of digital collections and provide search and retrieval service. Greenstone3 attempts to keep all the features of Greenstone2 with these changes to the collections and server. The following are major changes for the Greenstone3 runtime.

- Distributed Computing Support: Each Greenstone3 installation is a stand-alone system and can communicate with several sites on different computers. It provides a service that use XML (Extensible Markup Language) and SOAP (Simple Object Access Protocol) to delivery collection information to other computers.

- Interface Customisation: Greenstone3 uses XML and XSLT (Extensible Stylesheet Language Transformations) instead of "macros" for customising the visual display. The system generates data in XML and uses XSLT to convert to HTML (HyperText Markup Language). This approach should make it easier for end-users to create a customised look-and-feel for their collections.

- Cross Collection Search: The original Greenstone2 requires end-users to use the same index in order to search across all collections. In Greenstone3, the default index is used for each collection. Therefore, end-users can search all collections directly without creating same indexes.

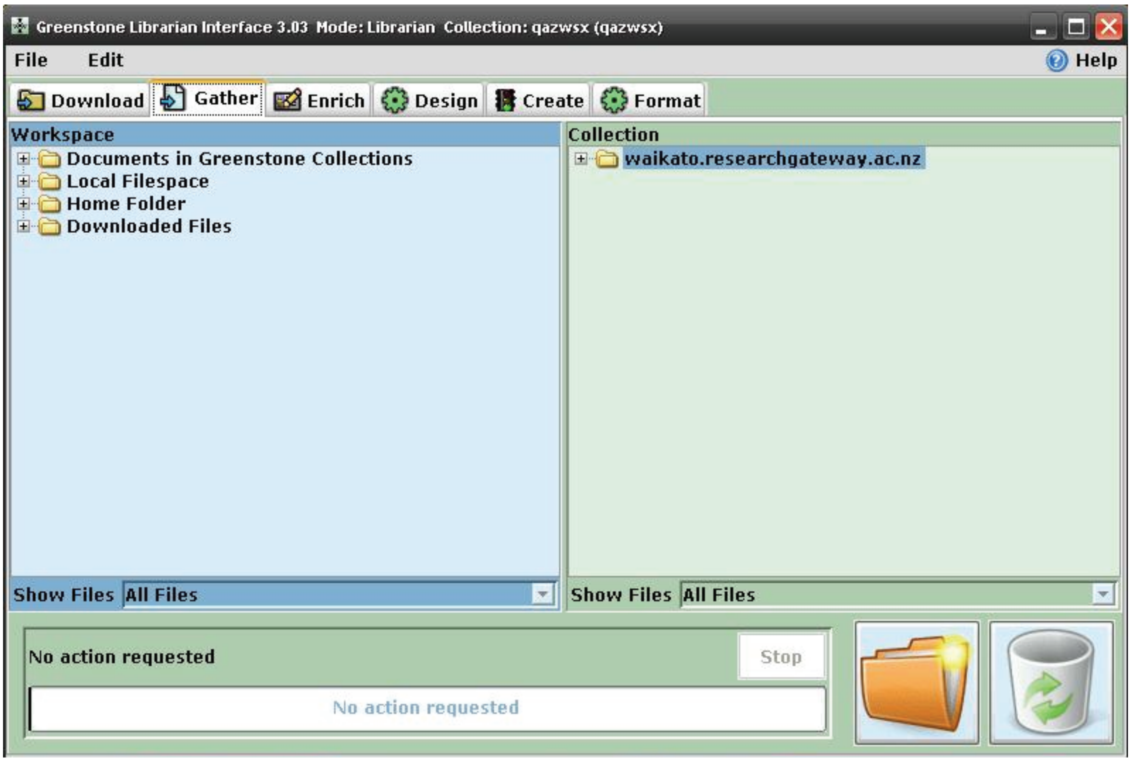
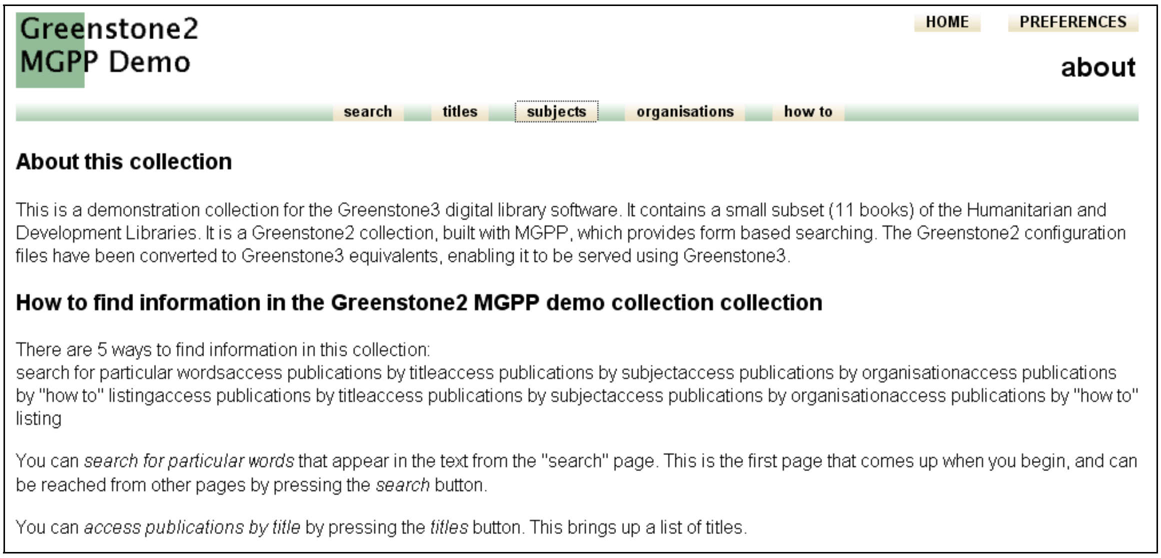


Figure 4 : Greenstone's Graphical Librarian Interface (GLI).



About this collection

This is a demonstration collection for the Greenstone3 digital library software. It contains a small subset (11 books) of the Humanitarian and Development Libraries. It is a Greenstone2 collection, built with MGPP, which provides form based searching. The Greenstone2 configuration files have been converted to Greenstone3 equivalents, enabling it to be served using Greenstone3.

How to find information in the Greenstone2 MGPP demo collection collection

There are 5 ways to find information in this collection:
search for particular wordsaccess publications by titleaccess publications by subjectaccess publications by organisationaccess publications by "how to" listingaccess publications by titleaccess publications by subjectaccess publications by organisationaccess publications by "how to" listing

You can *search for particular words* that appear in the text from the "search" page. This is the first page that comes up when you begin, and can be reached from other pages by pressing the *search* button.

You can *access publications by title* by pressing the *titles* button. This brings up a list of titles.

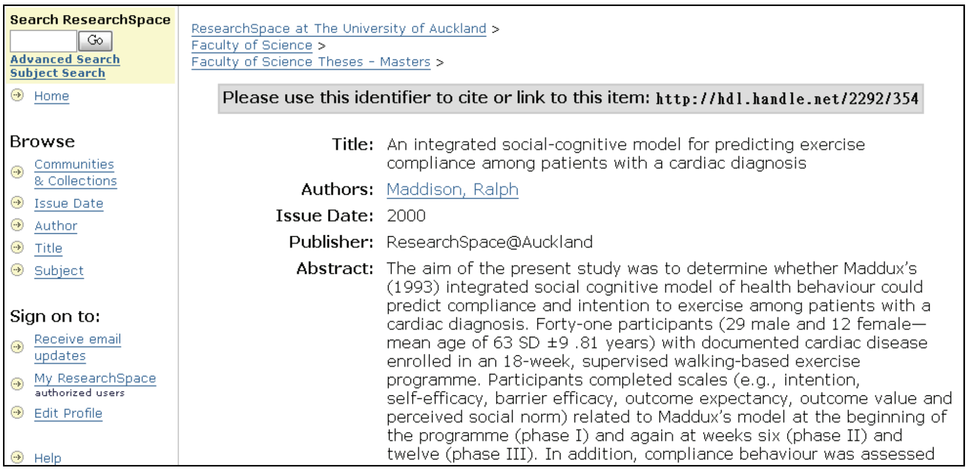
Figure 5 : Greenstone browsing interface

Greenstone website: http://www.greenstone.org/

2.2.2 DSpace

DSpace (http://www.dspace.org/) is an open source software application which commonly used for organising digital content into institutional repositories. It was designed to collect and preserve different kinds of digital materials such as computer programs, multimedia, videos, and texts and support browsing and searching for end-users. It was developed by MIT and HP and first released in 2002. It is written in Java and JSP (JavaServer Pages) and uses the Java Servlet Framework within a web server. DSpace has been used widely from large university libraries to small research organisations. DSpace uses Dublin Core Metadata to describe documents and mandates repository managers to use some of the metadata elements.

Figure 6 is a screenshot of a record in DSpace repository. On the left of Figure 6, there are five options for end-users to find information. End-users can search a document by its issue date, author, title, and subject. The document information page usually contains several metadata items to describe the resource. As shown in Figure 6, there are five metadata elements describing a document. DSpace website: http://www.dspace.org/



Please use this identifier to cite or link to this item: http://hdl.handle.net/2292/354

Title: An integrated social-cognitive model for predicting exercise compliance among patients with a cardiac diagnosis

Authors: Maddison, Ralph

Issue Date: 2000

Publisher: ResearchSpace@Auckland

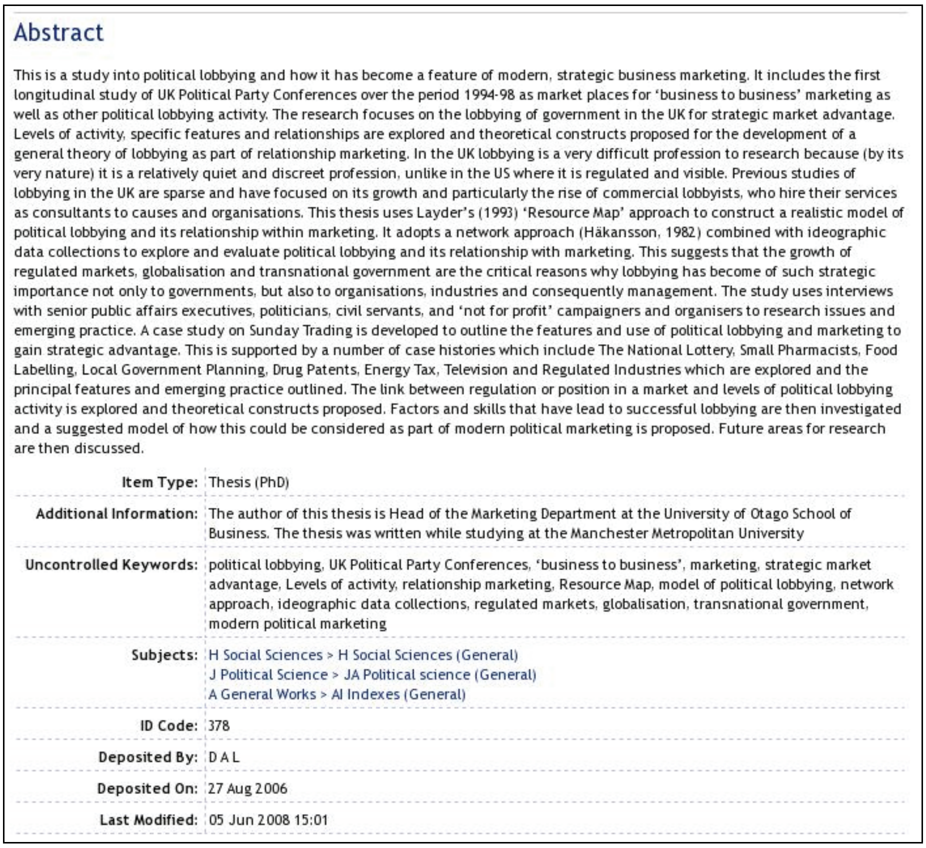
Abstract: The aim of the present study was to determine whether Maddux's (1993) integrated social cognitive model of health behaviour could predict compliance and intention to exercise among patients with a cardiac diagnosis. Forty-one participants (29 male and 12 female—mean age of 63 SD ±9 .81 years) with documented cardiac disease enrolled in an 18-week, supervised walking-based exercise programme. Participants completed scales (e.g., intention, self-efficacy, barrier efficacy, outcome expectancy, outcome value and perceived social norm) related to Maddux's model at the beginning of the programme (phase I) and again at weeks six (phase II) and twelve (phase III). In addition, compliance behaviour was assessed

Figure 6 : A screenshot of DSpace Repository.

2.2.3 EPrints

EPrints (http://www.eprints.org/) is an open source application for building high quality digital repositories and mainly used for institutional repositories. EPrints was first released in 2000 and one of the most widely used free open access, institutional repository software. EPrints is designed to preserve literature, scientific data, theses and papers from collections, exhibitions and performances. It encourages creators or organisations to digitalize their work and build digital repositories. This project is developed at the School of Electronics and Computer Science, University of Southampton, UK. Users can upload their documents via a simple and powerful web interface. Once the repository becomes a data provider, its digital materials can be shared on the Internet. It supports any kind of metadata schema and allows users to define the hierarchical structures for searching and viewing documents. EPrints can store any kind of digital material and check the completeness of the metadata automatically. EPrints is a Web and command-line application and written in Perl. It works under any UNIX platform but not Microsoft Windows operating system. Most institutions use DSpace and EPrints to build their digital repositories.

Figure 7 shows a part of a document information page. This information page contains a brief abstraction and uses eight metadata elements to describe the resource. The abstraction is extracted from the original document and allows users to preview it without downing the documents. EPrints website: http://www.eprints.org/software/



Abstract

This is a study into political lobbying and how it has become a feature of modern, strategic business marketing. It includes the first longitudinal study of UK Political Party Conferences over the period 1994-98 as market places for ‘business to business’ marketing as well as other political lobbying activity. The research focuses on the lobbying of government in the UK for strategic market advantage. Levels of activity, specific features and relationships are explored and theoretical constructs proposed for the development of a general theory of lobbying as part of relationship marketing. In the UK lobbying is a very difficult profession to research because (by its very nature) it is a relatively quiet and discreet profession, unlike in the US where it is regulated and visible. Previous studies of lobbying in the UK are sparse and have focused on its growth and particularly the rise of commercial lobbyists, who hire their services as consultants to causes and organisations. This thesis uses Layder’s (1993) ‘Resource Map’ approach to construct a realistic model of political lobbying and its relationship within marketing. It adopts a network approach (Häkansson, 1982) combined with ideographic data collections to explore and evaluate political lobbying and its relationship with marketing. This suggests that the growth of regulated markets, globalisation and transnational government are the critical reasons why lobbying has become of such strategic importance not only to governments, but also to organisations, industries and consequently management. The study uses interviews with senior public affairs executives, politicians, civil servants, and ‘not for profit’ campaigners and organisers to research issues and emerging practice. A case study on Sunday Trading is developed to outline the features and use of political lobbying and marketing to gain strategic advantage. This is supported by a number of case histories which include The National Lottery, Small Pharmacists, Food Labelling, Local Government Planning, Drug Patents, Energy Tax, Television and Regulated Industries which are explored and the principal features and emerging practice outlined. The link between regulation or position in a market and levels of political lobbying activity is explored and theoretical constructs proposed. Factors and skills that have lead to successful lobbying are then investigated and a suggested model of how this could be considered as part of modern political marketing is proposed. Future areas for research are then discussed.

Item Type:	Thesis (PhD)
Additional Information:	The author of this thesis is Head of the Marketing Department at the University of Otago School of Business. The thesis was written while studying at the Manchester Metropolitan University
Uncontrolled Keywords:	political lobbying, UK Political Party Conferences, ‘business to business’, marketing, strategic market advantage, Levels of activity, relationship marketing, Resource Map, model of political lobbying, network approach, ideographic data collections, regulated markets, globalisation, transnational government, modern political marketing
Subjects:	H Social Sciences > H Social Sciences (General) J Political Science > JA Political science (General) A General Works > AI Indexes (General)
ID Code:	378
Deposited By:	D A L
Deposited On:	27 Aug 2006
Last Modified:	05 Jun 2008 15:01

Figure 7 : Screenshot of a document in EPrints

2.3 Dublin Core

“The Dublin Core metadata standard is a simple yet effective element set for describing a wide range of networked resources” (Hillmann, 2005). The purpose of Dublin Core Metadata Set is to create a standard approach that can describe digital documents and makes them easy be retrieved. It is widely used to describe video, text, and image around the world. “The Dublin Core Metadata Element Set can be used for cross-domain resource description and resource discovery” (Dublin Core Metadata Initiative, 2008). Originally, it was designed for creators and authors to describe their works on the Internet. There are many libraries and museums using the Dublin Core metadata set to describe their collections at present. It includes two subsets: Simple Dublin Core Metadata Set and Qualified

Dublin Core Metadata Set. The simple Dublin Core Metadata Set is made up of fifteen elements (see Table 1); the qualified Dublin Core Metadata Set is created based on the simple Dublin Core with several additional elements to narrow the meanings of elements. There are many projects are using Dublin Core to build indexing systems for end-users to search and retrieve digital materials.

2.3.1 Simple Dublin Core

"Simple Dublin Core is a set of 15 metadata elements that represent a core set of elements likely to be useful across a broad range of vertical industries and disciplines of study" (Dublin Core Metadata Initiative, 2008). Every Dublin Core element is optional and may be used more than once if required. Table 1 shows the metadata standard. These fifteen elements can be extended by adding some parameters.

Simple Dublin Core does not mandate repository managers to use any controlled vocabulary. Two repository managers might use very different descriptions to describe the same resource. Different local policies for Dublin Core Metadata Element Set could also affect the use of this scheme. For example, repository A uses "format" elements to describe the size of files but repository B uses them to record the physical medium of the resources. Both repositories follow the rules but use them very differently.

Term Name	**Definition**
Title	A name given to the resource.
Creator	Examples of a Creator include a person, an organization, or a service.
Identifier	An unambiguous reference to the resource within a given context.
Subject	The topic of the resource.
Publisher	An entity responsible for making the resource available.
Description	An account of the resource.
Contributor	An entity responsible for making contributions to the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Source	A related resource from which the described resource is derived.
Language	A language of the resource.
Relation	A related resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Rights.	Information about rights held in and over the resource.
Type	The nature or genre of the resource.

Table 1 : Dublin Core Metadata Element Set

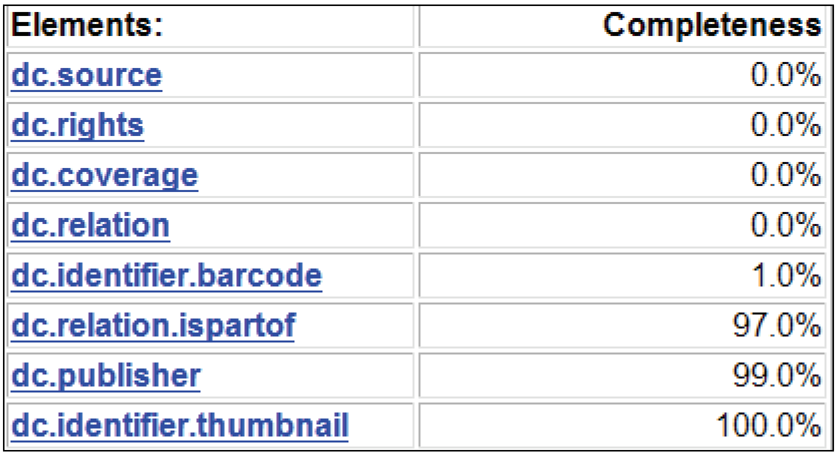
(http://dublincore.org/documents/dces/)

2.3.2 Qualified Dublin Core

After the specification of the original 15 elements was defined, the Dublin Core Metadata Initiative started to refine the Dublin Core Metadata Element Set. Some of the elements are ambiguous and can be redefined. For instance, Date can be used to express creation date, last modified date, or issue date. Therefore, the qualified Dublin Core is design to reduce ambiguity of the simple Dublin Core.

> *"Qualified Dublin Core consists of the 15 elements from Simple Dublin Core, along with additional elements, element refinements, vocabulary encoding schemes, and syntax encoding schemes. A refined element shares the meaning of the unqualified element but with a more restricted scope" (Dublin Core Metadata Initiative, 2008).*

The qualified Dublin Core Metadata Set makes the meaning of an element more precise and specific. Repository managers could use "dc.Date.issue" element to indicate the issue date rather than "dc.Date" element. Figure 8 shows that how "dc.Identifier" element can be extended by adding new attributes.



Elements:	Completeness
dc.source	0.0%
dc.rights	0.0%
dc.coverage	0.0%
dc.relation	0.0%
dc.identifier.barcode	1.0%
dc.relation.ispartof	97.0%
dc.publisher	99.0%
dc.identifier.thumbnail	100.0%

Figure 8 : Screenshot of a Dublin Core Metadata Set with qualified metadata elements

2.4 Open Archives Initiative

> *"The Open Archives Initiative (OAI) is an attempt to build a low-barrier interoperability framework for archives (institutional repositories) containing digital content (digital libraries)" (Lagoze et al., 2001).*

The Open Archives Initiative (OAI) was designed to build a standard for exchanging archives containing digital content and developed by DLF (Digital Library Federation) (Coalition for Networked Information--CNI) NSFG (National Science Foundation Grant).

In 2000, the OAI project extends its scope to the digital libraries to enhance the interoperability framework. The Open Archives Initiative for Metadata Harvesting (OAI-PMH) (Lagoze et al., 2002) is a protocol that facilitates digital library interoperability and cross-domain resource discovery. It defines a mechanism for data providers to expose their metadata and mandates them to map their metadata to the unqualified Dublin Core (i.e. Simple Dublin Core). The OAI-PMH requires XML-encoded metadata to support harvesting and reuse by other data providers. OAI technical committee developed the OAI-PMH from Kahn-Wilensky framework (Kahn and Wilensky, 1995) and Dienst open architecture (Davis and Lagoze, 2000). The OAI-PMH uses HTTP to transfer metadata records and the metadata records are stored as XML files. The OAI-PMH can be implemented in Perl, Java, C++, and other programming languages. As mentioned earlier, the OAI was developed to increase interoperability standards specifically for enhancing access to e-print archives initially. However, this technology and standard have been used in a much broader domain. It has been expanded to promote broad access to digital resources for eScholarship, eLearning, and eScience. Figure 9 explains how service providers harvest metadata via OAI-PMH. Firstly, service providers send requests to data providers. Then data providers generate a XML encoded message and return it to service providers. Table 2 gives a brief summary of the meaning of the OAI verbs.

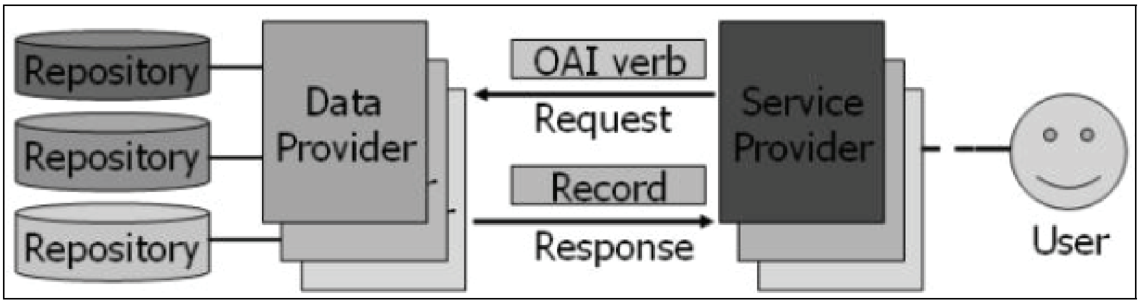


Figure 9 : How the OAI-PMH works

(http://www.lib.tku.edu.tw/esource/scholar/project/OAI.htm)

OAI Verb	Definition
GetRecord	This verb is used to retrieve an individual metadata record from a repository.
Identify	This verb is used to retrieve information about a repository.
ListIdentifiers	This verb is an abbreviated form of ListRecords, retrieving only headers rather than records.
ListMetadataFormats	This verb is used to retrieve the metadata formats available from a repository.
ListRecords	This verb is used to harvest records from a repository.
ListSets	This verb is used to retrieve the set structure of a repository.

Table 2 : Definitions of OAI verbs (Lagoze et al., 2002)

2.5 Metadata

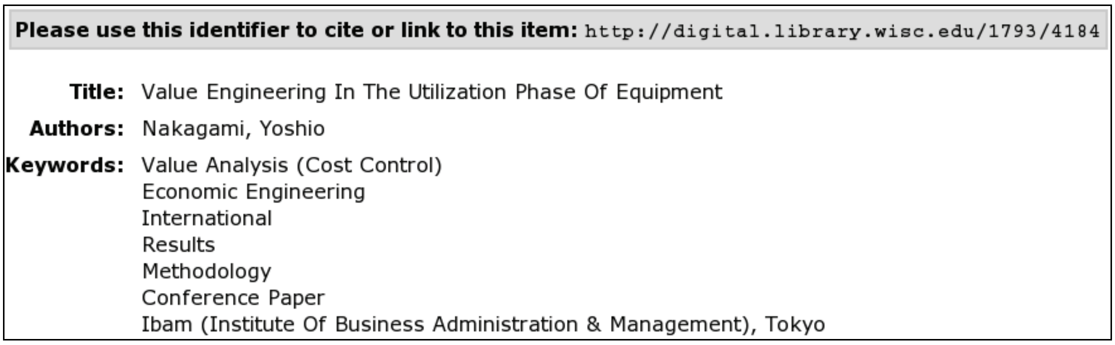
> *"Data quality is important in the digital libraries because high quality data insures accurate and complete access to online objects" (Beall, 2005).*

Beall (2005) explains there are two aspects of digital library data quality: quality of data in the objects, and the quality of the metadata associated with the objects. In this research project, we focus on the metadata quality rather than the quality of data.

There are always some inaccurate, inconsistent, and incomplete metadata existed in the digital repositories. One study counted the number of "visible" errors in each record (e.g., spelling or typographical errors, file formatting errors, or incorrect date formats); in the sample, 10-30% of records featured such errors and 30% of records contain blank (labelled but null value) elements (Moen et al., 1997, Efron, 2007). Beall and Kafadar (2007) show that retrieval effectiveness can be significantly affected by typographical errors, with the results heavily dependent on the particularly search terms used. Therefore, it is necessary to ensure digital repositories have good quality metadata in order to access to resources. In a small repository, the metadata quality can be controlled by checking metadata records manually. As the number of records increases, manual quality checks become infeasible.

Figure 10 and Figure 11 are screenshots of two different resources in a same repository. Each resource is described by several keywords that help end-users to locate the item. As shown in Figure11, "Administration" is misspelled as

"Adminstration". Fortunately, end-users still can find it by using other keywords. If this mistake appears in the title, it will block access to the resource.



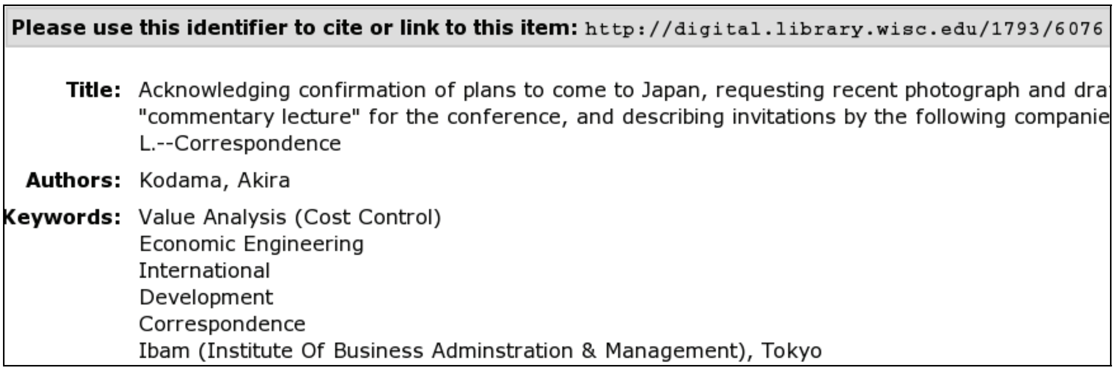
Please use this identifier to cite or link to this item: http://digital.library.wisc.edu/1793/4184

Title: Value Engineering In The Utilization Phase Of Equipment

Authors: Nakagami, Yoshio

Keywords: Value Analysis (Cost Control)
Economic Engineering
International
Results
Methodology
Conference Paper
Ibam (Institute Of Business Administration & Management), Tokyo

Figure 10 : A screenshot of a metadata record with no spelling errors



Please use this identifier to cite or link to this item: http://digital.library.wisc.edu/1793/6076

Title: Acknowledging confirmation of plans to come to Japan, requesting recent photograph and dra
"commentary lecture" for the conference, and describing invitations by the following companie
L.--Correspondence

Authors: Kodama, Akira

eywords: Value Analysis (Cost Control)
Economic Engineering
International
Development
Correspondence
Ibam (Institute Of Business Adminstration & Management), Tokyo

Figure 11 : A screenshot of a metadata record with a spelling error -

Administration is misspelled as "Adminstration".

There are several studies discussing the impacts of poor metadata quality and how metadata quality can be measured. The initial efforts in metadata development have been primary invested in the structure rather than in content (Bruce and Hillmann, 2004). However the metadata quality does not only depend on the standards but also the content (Duval and Ochoa, 2006). The two main approaches can be found in research field of metadata quality: Statistical Data (Barton et al., 2003; Bruce and Hillmann, 2004; Guy et al., 2004, Moen and McCluren, 1997; Najjar et al., 2003; Zeng, et al., 2004) and Visualisation (Hillmann and Dushay, 2003). In the following sections, we give an overview of information and metadata quality assessment.

2.5.1 Metadata Quality

“The term "meta" comes from a Greek word that denotes something transcendental or beyond nature” (Dublin Core Metadata Initiative, 2008). It can be referred to the descriptive information of the resources. “A definition that can be used is: high quality metadata supports the functional requirements of the system it is designed to support, which can be summarised as quality is about fitness for purpose” (Guy et al., 2004). A metadata record consists of a set of attributes, or elements to describe the resource. For example, metadata records with elements that describe a book such as author, title, and subject.

In this research, we will consider the “metadata quality as the measure of fitness for a task” (Duval and Ochoa, 2006).The quality of metadata could affect the use of digital repositories. Digital repository managers need to maintain a high level of consistency and quality of metadata in order to benefit end-users.

2.5.2 Statistics Data

Duval and Ochoa (2006) suggest two main approaches to metadata quality evaluation:

- *Manual Quality Evaluation*. Some studies try to analyse metadata quality problems by reviewing a statistical significant sample of metadata records and comparing the values with those generated by experts manually. However, “the metadata quality estimation is only valid for the whole repository at a given point in time” (Duval and Ochoa, 2006). Once new records have been added into the repository, the metadata quality estimation is no longer accurate. Human experts have to estimate the repository again to obtain the latest metadata statistics. Therefore, this approach is only suitable for small-size repositories (Barton et al., 2003; Bruce and Hillmann, 2004; Guy et al., 2004). Table 3 shows a summary of different quality evaluation studies.

□ *Statistical Quality Evaluation*. The approach used by (Moen and McCluren, 1997; Najjar et al., 2003; Zeng et al., 2004) collects the statistical information from all metadata records and estimates the usage of them. This approach can provide some statistical information of collections but the information is not clear enough to indicate the metadata quality. For example, statistical information shows that the overall completeness of a collection is 55%. This report would be able to show repository managers which metadata elements are not complete. However, this report is not very useful for finding individual errors. Hughes (2004) uses a similar approach to calculate the completeness at the repository level for each of the repository in the Open Language Archive.

□

Study	**Approach**	**No. of Records**	**Main focus of evaluation**
Greenberg et al. (2001)	Manual	11	Quality of non-expert metadata
Stivila et al. (2004)	Manual	150	Identify quality of records
Wilson (2007)	Manual	100	Quality of non-expert metadata
Hughes (2004)	Statistical	27,000	Completeness of records
Najjar et al. (2004)	Statistical	3,700	Usage of the metadata standard
Bui and Park (2006)	Statistical	1,040,034	Completeness of records

Table 3 : Review of different quality evaluation studies (Duval and Ochoa, 2006)

"There is a wide range agreement on the need to have high quality metadata but less consensus on what high quality metadata means and much less in how it should be measured" (Duval and Ochoa, 2006). Moen et al. (1998) identifies 23 quality parameters but some of them are more focused on the metadata standards. Gasser and Stvilia (2001) define another metadata quality framework by analysing 32 representative quality assessment frameworks and group them in three dimensions: Intrinsic IQ, Relational/Contextual IQ and Reputational IQ. The framework defined by Gasser and Stvilia is intended to be general enough to apply to different kinds of information.

- *Intrinsic IQ*: The Intrinsic Information Quality dimension assesses information by measuring its metadata attributes rather than its context. The Intrinsic IQ dimension does not depend much on context but attributes. Hence it can be measured more objectively.

- *Relational/Contextual IQ*: The Relational/Contextual Information Quality dimension focuses on the relationships between the information and its context. It measures how well an information object describes the context. Since the related object can change independently, the relational/contextual IQ dimensions of an information item are not unchangeable.

- *Reputational IQ*: The Reputational Information Quality dimension measures the position or reputation of the information in the community. It is often determined by its origin and its record of information.

Bruce and Hillmann (2004) define an examination of characteristics of metadata quality based on Gasser and Stvilla framework to improve its applicability. The categorization of quality measures defined by Bruce and Hillmann is part of a Quality Assurance Framework (QAF) developed by Statistics Canada (STC). "The original STC QAF described six dimensions of information quality: relevance, accuracy, timeliness, accessibility, interpretability, and coherence" (Bruce and Hillmann, 2004). Bruce and Hillmann devise seven criteria for determining metadata quality:

- *Completeness*: "Completeness is the degree to which the metadata records contents all the information needed to have an ideal representation of the described object" (Duval and Ochoa, 2006). The completeness quality value can be measured by different approaches. The main approach is counting the number of fields that are not empty to assess the completeness of metadata records. In the case of multi-valued fields, the metadata element is considered complete if at least one field is filled. However, this approach does not reflect the quality very well because not all metadata elements are equally important. Therefore, this metric can be modified by adding the weight values to the metadata elements. It implies a difference-weighted completeness value could be calculated for different contexts. For example, titles may be more important than formats and contributors for end-users.

- *Accuracy*: "The accuracy is the degree to which the metadata elements match the objects" (Duval and Ochoa, 2006). If an item is described correctly by its metadata, the metadata accuracy is high. For objective information such as the file format or language, it is easy to know whether an item is described correctly or not. But in the case of subjective information such as title, author's name, it is more complex to check. Human experts can assess metadata accuracy easy but computers require complex algorithm and lots of resource to simulate human intelligence. "Accuracy is simply high-quality editing: the elimination of typographical errors, conforming expression of person names and place names, use of standard abbreviation, and so on" (Bruce and Hillmann, 2004). A typographical error is considered as a part of the accuracy dimension.

- *Provenance*: Provenance is a useful measurement for judging metadata quality (Bruce and Hillmann, 2004). "Provenance is a statement of any changes in the ownership and custody of the resource that are significant for its authenticity and integrity" (Dublin Core Collection Description Working Group, 2004). It could help repository managers judge the credibility of the target objects. If the provenance shows that the target object has very low credibility, repository managers might remove it from the collection.

- *Conformance to expectations*: The conformance to expectations can be explained as metadata elements fulfil the requirements of users. Metadata element sets should contain elements that users would expect to use and find. They should not contain any irrelevant and unnecessary information.

- *Logical consistency, coherence*: Collections do not exist in isolation. Hence, repository managers need to ensure that metadata elements are consistent with standard definition and similar objects (Bruce and Hillmann, 2004). "The logical consistency and coherence is the degree to which a metadata record matches a standard definition and the values used in the fields correlate positively among them" (Duval and Ochoa, 2006). If similar objects contain consistent metadata values, end-users will be able to use similar criteria to access them.

- *Timeliness*: Timeliness can be interpreted as "currency" and "lag". The currency problems happen when object content changes but the metadata does not. Metadata is out of date if it loses the synchronization with its target object. The lag can be measured as the interval between the released date of the target object and the point at which the metadata becomes knowable or available (Bruce and Hillmann, 2004).

- *Accessibility*: "The accessibility measures the degree to which metadata is accessible both in terms of cognitive accessibility as well as physical/logical accessibility" (Bruce and Hillmann, 2004; Duval and Ochoa, 2006). The cognitive accessibility measures how easy it can be understood by the users. The physical/logical accessibility could be understood as how easy is to find records in the repository.

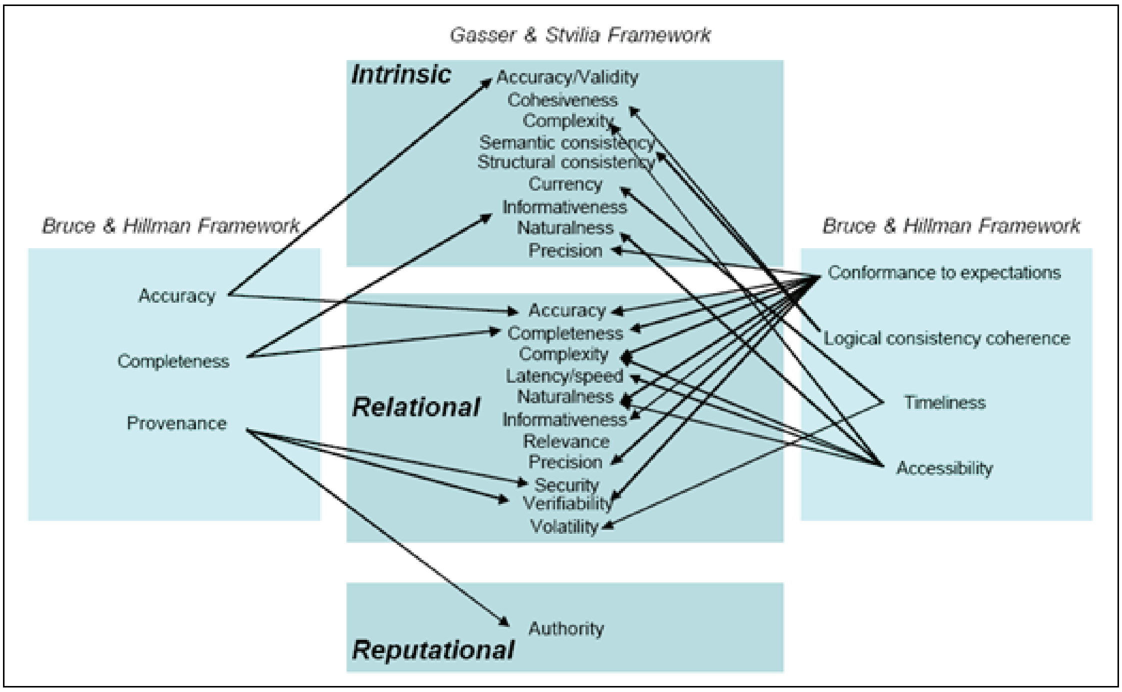


Figure 12 : Mapping between the Bruce & Hillmann and the Gasser & Stvilia frameworks, from Shreeves et al. (2005)

Although the metadata quality frameworks are defined by two different groups, the basic ideas are very similar. It is clear that the Gasser and Stvilia metadata quality framework can be interpreted as a broader version of Bruce and Hillmann's concept in Figure 12.

The Bruce and Hillmann metadata quality framework defines several formulas and these calculations can be done automatically by computer software. The metrics can be used for a wide range of digital repositories as digital libraries or museum catalogs. However, the statistical data aggregated at the repository level may not be able to indicate repository managers which metadata field is missing, which metadata value is incorrect.

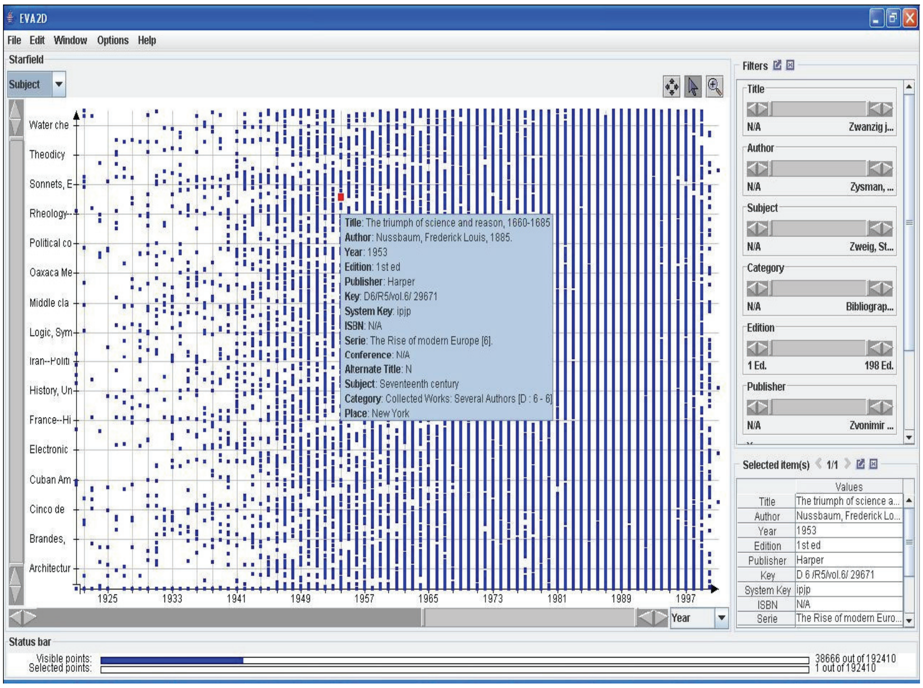
2.5.3 Visualisation Tools

Statistical data can be used to create visualisations of a repository in order to gain better understanding of the distribution of quality problems. "The use of data visualisation software can significantly improve efficiency and thoroughness of metadata evaluation" (Hillmann and Dushay, 2003). Visualisation tools usually allow end-users to access more details at the document and metadata element levels.

Starfield Displays (Ahlberg and Shneiderman, 1994) are the well known visualisation tools. Starfield was first introduced at the University of Maryland's Human-Computer Interaction Lab. A Starfield display transforms data to a two dimensional grid and use small dots to represent metadata elements. In Figure 13, the horizontal axis represents years while the vertical axis represents subject categories. There are scroll bars next to the axes for user to adjust the range of values. Starfield has a filtering mechanism and allows users to view the certain rage of data. The scroll bars on the right hand side represent the item attributes and allow users to adjust values and create new visual displays. This filtering mechanism can be useful for repository managers to explore metadata elements in large repositories (Sánchez et al., 2007; Sánchez et al., 2005).

When the cursor is pointing to a dot, users will see the details of that metadata element. It allows users to select an area and magnify it to view properly. Due to the numbers of items in the collections, Starfield uses a technology call "cluttered representation" to display the visualisation (Sánchez et al., 2007). For a large repository, it is almost impossible to use a single dot to represent an item. In this case, this tool uses a dot to represent multiple items. Figure 14 illustrates the case when users use the zoom option or the filtering mechanisms to locate a specific item. The intervals between each value become wider and the data is more clearly to see for users. Users can easily identify metadata problems such as spelling errors and incorrect values if they use visualisation tools.

Figure 15 shows an example of error in "date" element. It is clear there is a cluster of dots located on the y-axis. This indicates that those items were published in 0AD (Sánchez et al., 2007). Apparently, it is incorrect and those metadata elements should be revised. If metadata elements do not have values, they will not appear on the plot.

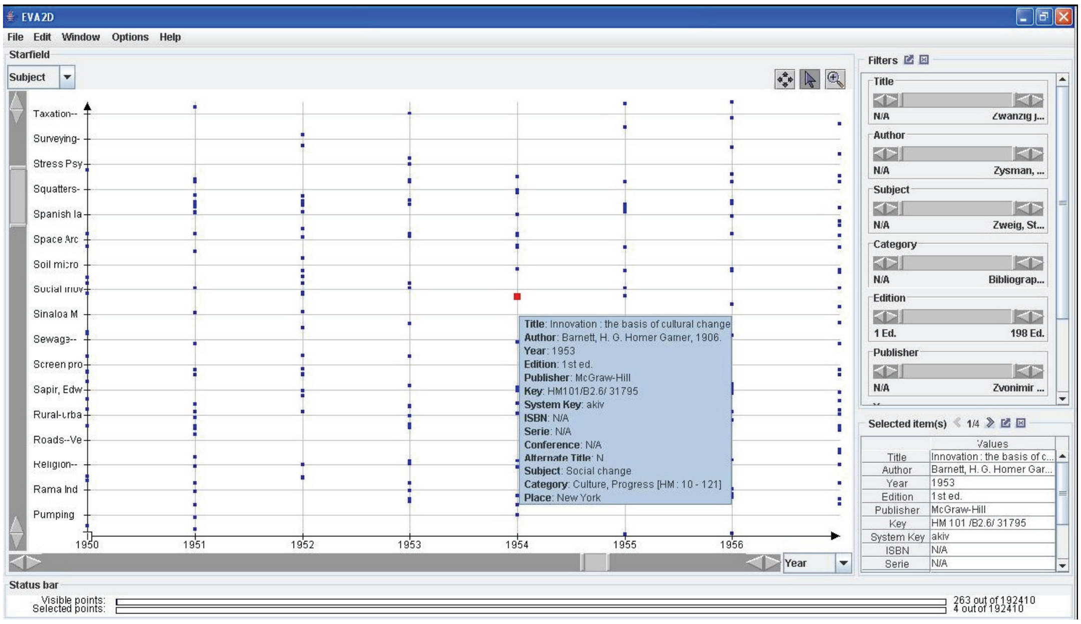


Title: The triumph of science and reason, 1660-1685
Author: Nussbaum, Frederick Louis, 1885.
Year: 1953
Edition: 1st ed
Publisher: Harper
Key: D6/R5/vol.6/ 29671
System Key: ipjp
ISBN: N/A
Serie: The Rise of modern Europe [6].
Conference: N/A
Alternate Title: N
Subject: Seventeenth century
Category: Collected Works: Several Authors [D : 6 - 6]
Place: New York

	Values
Title	The triumph of science a...
Author	Nussbaum, Frederick Lo...
Year	1953
Edition	1st ed
Publisher	Harper
Key	D 6 /R5/vol.6/ 29671
System Key	ipjp
ISBN	N/A
Serie	The Rise of modern Euro...

Visible points: 38666 out of 192410
Selected points: 1 out of 192410

Figure 13 : Starfield-based interface for an online library catalog from Sánchez et al. (2007)

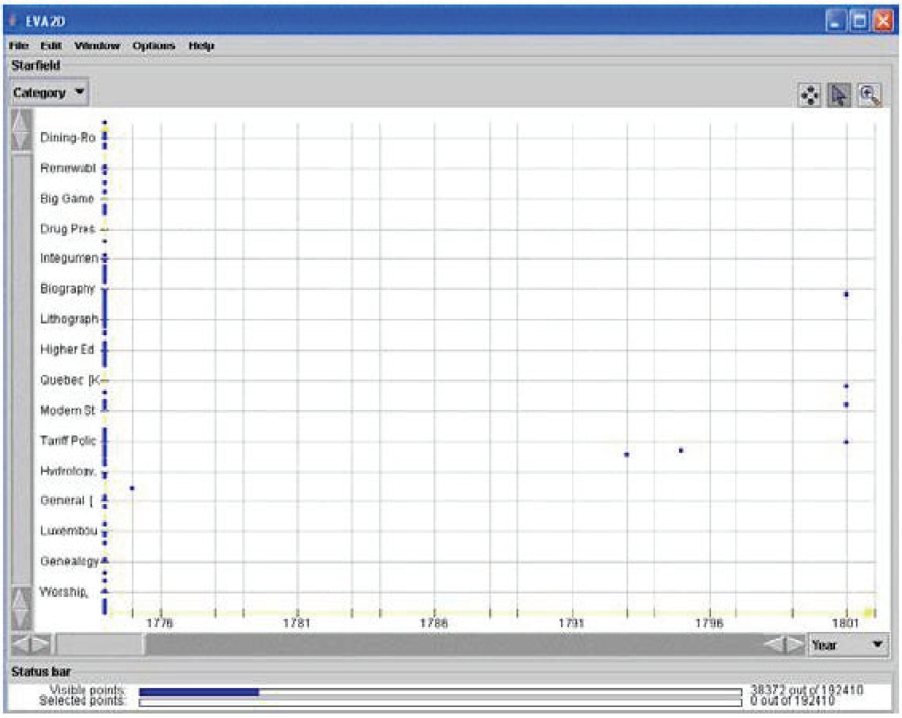


Title: Innovation : the basis of cultural change
Author: Barnett, H. G. Homer Garner, 1906.
Year: 1953
Edition: 1st ed.
Publisher: McGraw-Hill
Key: HM101/B2.6/ 31795
System Key: akiv
ISBN: N/A
Serie: N/A
Conference: N/A
Alternate Title: N
Subject: Social change
Category: Culture, Progress [HM : 10 - 121]
Place: New York

	Values
Title	Innovation : the basis of c...
Author	Barnett, H. G. Homer Gar...
Year	1953
Edition	1st ed.
Publisher	McGraw-Hill
Key	HM 101 /B2.6/ 31795
System Key	akiv
ISBN	N/A
Serie	N/A

Visible points: 263 out of 192410
Selected points: 4 out of 192410

Figure 14 : Starfield after zooming-in operation from Sánchez et al. (2007)



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

Figure 15 : Metadata errors in data from Sánchez et al. (2007).

Spotfire (http://spotfire.tibco.com/index.cfm) is the commercial version of Starfield and it is a visual graphical analysis application that allows users to browse and analyse metadata element values at the same time. Spotfire has been specially developed for quality analysis of repository data (Hillmann and Dushay, 2003). It was launched in mid-1996 by IVEE Development, which became renamed as TIBCO.

Spotfire provides six data dimensions for users to explore metadata elements and can display large quantities of data on the screen like Starfield. Spotfire user interface also provides zoom in and zoom out option on any portion of the data and is capable of full text searching. This visual graphical tool allows users to review large quantities of data efficiently. It also allows users to select data based on relevant characteristics such as “don’t display empty element” or “look for all values starts with http://” (Hillmann and Dushay, 2003).

Figure 16 presents a view of the overall structure of a collection’s metadata elements. The horizontal axis represents document identifiers while the vertical axis represents metadata elements. Spotfire uses different colours and size of boxes to represent each metadata data. Repository managers could easily discover which metadata elements are not fully used by looking at the patterns of elements and fix the incorrect or missing data. Figure 16 shows almost every document uses at least one “language” element but only small number of them contains “alternative” elements. “A Spotfire plot allows users to detect patterns: the presence or absence of fields in a collection’s metadata patterns with particular fields or within groups of records" (Hillmann and Dushay, 2003).

Spotfire also provides a table view of the data, which is similar to a spreadsheet. Table views are more useful than the statistical data described earlier for detecting errors. Figure 17 is a screenshot of Spotfire's table view for a data field. All metadata values could be examined at the same time in the table view as show in Figure 17. If there are any typographical errors or incorrect data in the metadata values, the sorted table views could assist repository managers to detect them.

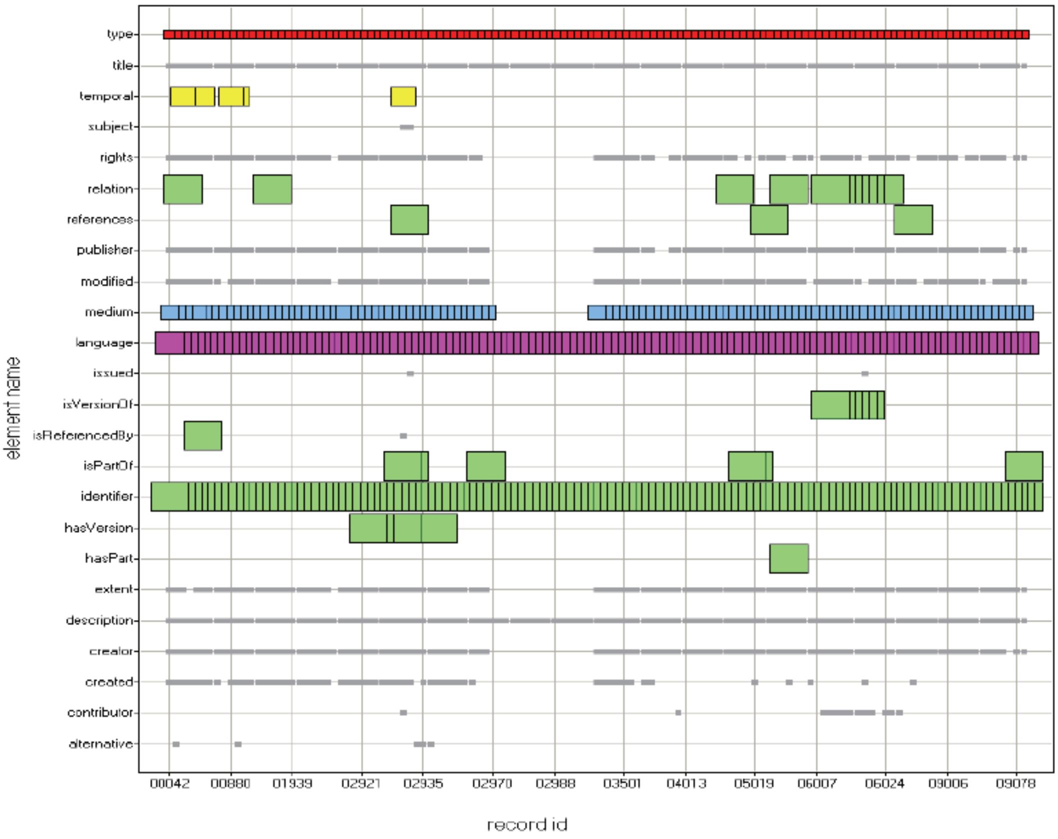
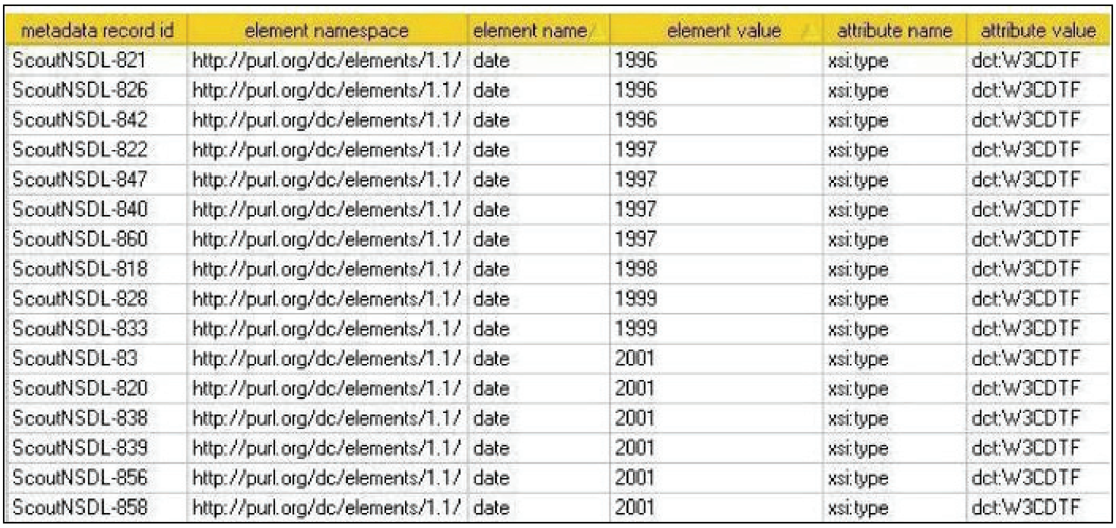


Figure 16 : Spotfire scatter plot for a collection's metadata from Hillmann and Dushay (2003)



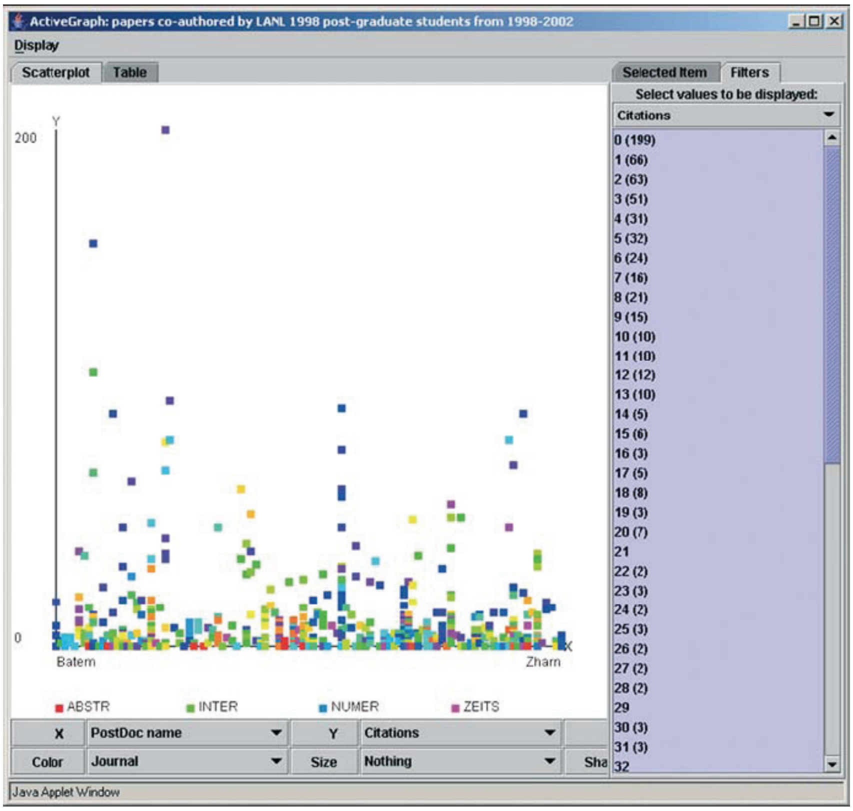
metadata record id	element namespace	element name	element value	attribute name	attribute value
ScoutNSDL-821	http://purl.org/dc/elements/1.1/	date	1996	xsi:type	dct:W3CDTF
ScoutNSDL-826	http://purl.org/dc/elements/1.1/	date	1996	xsi:type	dct:W3CDTF
ScoutNSDL-842	http://purl.org/dc/elements/1.1/	date	1996	xsi:type	dct:W3CDTF
ScoutNSDL-822	http://purl.org/dc/elements/1.1/	date	1997	xsi:type	dct:W3CDTF
ScoutNSDL-847	http://purl.org/dc/elements/1.1/	date	1997	xsi:type	dct:W3CDTF
ScoutNSDL-840	http://purl.org/dc/elements/1.1/	date	1997	xsi:type	dct:W3CDTF
ScoutNSDL-860	http://purl.org/dc/elements/1.1/	date	1997	xsi:type	dct:W3CDTF
ScoutNSDL-818	http://purl.org/dc/elements/1.1/	date	1998	xsi:type	dct:W3CDTF
ScoutNSDL-828	http://purl.org/dc/elements/1.1/	date	1999	xsi:type	dct:W3CDTF
ScoutNSDL-833	http://purl.org/dc/elements/1.1/	date	1999	xsi:type	dct:W3CDTF
ScoutNSDL-83	http://purl.org/dc/elements/1.1/	date	2001	xsi:type	dct:W3CDTF
ScoutNSDL-820	http://purl.org/dc/elements/1.1/	date	2001	xsi:type	dct:W3CDTF
ScoutNSDL-838	http://purl.org/dc/elements/1.1/	date	2001	xsi:type	dct:W3CDTF
ScoutNSDL-839	http://purl.org/dc/elements/1.1/	date	2001	xsi:type	dct:W3CDTF
ScoutNSDL-856	http://purl.org/dc/elements/1.1/	date	2001	xsi:type	dct:W3CDTF
ScoutNSDL-858	http://purl.org/dc/elements/1.1/	date	2001	xsi:type	dct:W3CDTF

Figure 17 : Spotfire table view for a data field from Hillmann and Dushay (2003)

ActiveGraph (Marks et al., 2005) is an information visualisation tool designed to provide users with a customisable view of objects in a digital library. ActiveGraph provides an efficient scatter plot of objects and allows users to analyse the data set. The data set can represent any digital library objects such as books, journals, papers, and images. "Since ActiveGraph is intended for use in the context of digital libraries, data attributes consist for the most part of metadata fields such as title, author, date of publication, and number of citations" (Marks et al., 2005). Figure 18 is a screenshot of the ActiveGraph scatter plot for the LANL digital library. The postdoc's name is mapped to the X-axis and the number of citations is mapped to the Y-axis. The scatter plot shows that one paper has been cited more than 200 times and many papers are not cited at all. Users can use the menus at the bottom to change the X-axis and Y-axis.

The most interesting feature of ActiveGraph is that users can view and analyse the set of objects without scrolling or paging. The filtering mechanism provided by the ActiveGraph is not like other visualisation tools. Its filtering mechanism provides a list of metadata values for users to select. Once a value is selected, the graphical interface will highlight that value on the scatter plots. Figure 19 is a screenshot of a scatter plot after the filtering mechanism has applied. The journal category filter is used and values “COMP” and “MATH” are selected. This tool displays all papers published by computer science and mathematics departments on the left hand side of the screen (Marks et al., 2005).

“ActiveGraph includes several features that are not related to information visualisation, but are important to retrieve and analysis in the context of digital libraries” (Marks et al., 2005). ActiveGraph allows users to view and customise the contents of a collection, edit metadata, add annotations and create new elements. These features accommodate the needs of different users. For example, end-users may want to add annotations to describe their works and repository managers can use this tool to create metadata elements.



ActiveGraph: papers co-authored by LANL 1998 post-graduate students from 1998-2002

Display

Scatterplot | Table

Selected Item | Filters

Select values to be displayed:

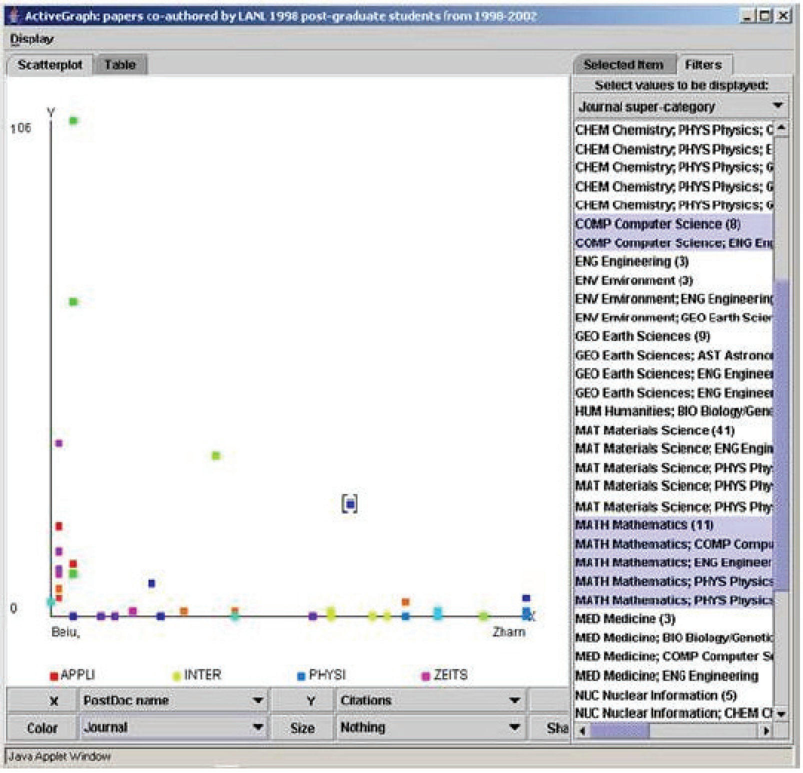
Citations

0 (199)
1 (66)
2 (63)
3 (51)
4 (31)
5 (32)
6 (24)
7 (16)
8 (21)
9 (15)
10 (10)
11 (10)
12 (12)
13 (10)
14 (5)
15 (6)
16 (3)
17 (5)
18 (8)
19 (3)
20 (7)
21
22 (2)
23 (3)
24 (2)
25 (3)
26 (2)
27 (2)
28 (2)
29
30 (3)
31 (3)
32

X: PostDoc name
Y: Citations
Color: Journal
Size: Nothing

Java Applet Window

Figure 18 : ActiveGraph scatter plot from Marks et al. (2005)



ActiveGraph: papers co-authored by LANL 1998 post-graduate students from 1998-2002

Display

Scatterplot | Table

Selected Item | Filters

Select values to be displayed:

Journal super-category

CHEM Chemistry; PHYS Physics; C
CHEM Chemistry; PHYS Physics; E
CHEM Chemistry; PHYS Physics; G
CHEM Chemistry; PHYS Physics; G
CHEM Chemistry; PHYS Physics; G
COMP Computer Science (8)
COMP Computer Science; ENG En
ENG Engineering (3)
ENV Environment (3)
ENV Environment; ENG Engineerin
ENV Environment; GEO Earth Scie
GEO Earth Sciences (9)
GEO Earth Sciences; AST Astrono
GEO Earth Sciences; ENG Enginee
GEO Earth Sciences; ENG Enginee
HUM Humanities; BIO Biology/Gen
MAT Materials Science (41)
MAT Materials Science; ENG Engin
MAT Materials Science; PHYS Phy
MAT Materials Science; PHYS Phy
MAT Materials Science; PHYS Phy
MATH Mathematics (11)
MATH Mathematics; COMP Compu
MATH Mathematics; ENG Enginee
MATH Mathematics; PHYS Physics
MATH Mathematics; PHYS Physics
MED Medicine (3)
MED Medicine; BIO Biology/Genetic
MED Medicine; COMP Computer S
MED Medicine; ENG Engineering
NUC Nuclear Information (5)
NUC Nuclear Information; CHEM C

X: PostDoc name
Y: Citations
Color: Journal
Size: Nothing

Java Applet Window

Figure 19 : ActiveGraph scatter plot after a filter has been applied from Marks et al. (2005)

2.3 Summary

A useful metadata quality tool for digital repositories should generate meaningful statistics and visualisations. At the beginning of this chapter, three different software systems for building digital libraries were introduced. Each of them plays an important role in this research. Greenstone can gather OAI metadata records and use them to build collections. DSpace and EPrints are popular software systems for building repositories to store digital materials such as video, images, and text.

There have been several studies that are related to quality of metadata in digital repositories. In section 2.5, the literature shows how people can use quality metrics and formulae to produce quality values for metadata. There are two main approaches: Manual Quality Evaluation and Statistical Quality Evaluation. Manual quality evaluation requires a human to review metadata records individually; statistical quality evaluation generates numerical information from all of the metadata records of a collection. A manual quality evaluation requires human experts; the statistics quality evaluation does not indicate specific metadata errors.

From the literature we conclude:

- Metadata quality is important for the development of digital collections.
- Manual approaches of measuring metadata quality are infeasible for large repositories.
- There are only small parts of information quality frameworks (Bruce and Hillmann, 2004; Gasser and Stvilia, 2001) that can be easily automated.

- There is little evidence and few experience reports that show metadata quality tools in use (Nichols and Chan et al., 2008).
- Visualisation tools can be potentially useful for repository managers to detect metadata errors.
- A useful metadata quality tool is likely to have both statistical and visualisation components
- More studies with authentic data are needed to clarify requirements for metadata quality tools (Nichols and Chan et al., 2008).

Due to the lack of evidence and experience reports, we decided to integrate some existing computer software applications to build a tool that could assist repository managers to maintain the metadata quality of collections and detect errors. Three main components are listed below:

1. Greenstone: Greenstone is used to gather OAI records from remote repositories and organise into collections.

2. DSpace / EPrints: Most institutional repositories use either DSpace or EPrints to build collections and these two software applications also support OAI-PMH.

3. OAI-PMH: OAI-PMH is a protocol for metadata harvesting. Institutional repositories can make available metadata via OAI-PMH and Greenstone can make OAI-PMH service requests to harvest that metadata.

The following chapter describes the development of a tool to address these issues.

Chapter 3

System Design and Implementation

3.1 Introduction

This chapter discusses the design and the implementation of a metadata quality tool. Section 3.2.1 gives a brief introduction of the system and user requirements. Section 3.2.2 summarises a series of use cases that were used to give guide development. Section 3.2.3 discusses the implementation rationale, overall structure of the system and issues that occurred during the implementation stage. In section 3.3, we discuss finding potential duplicates in metadata records.

We refer to the developed system as Mat, for *M*etadata *A*nalysis *T*ool.

3.2 Building a Metadata Quality Tool

In this section, we will discuss the basic activities of the requirements specification. A software requirements specification is a complete description of the behaviour of the system to be developed. There are two types of requirements specification: user requirements and system requirements. We will focus on user requirements in this section.

3.2.1 Introduction of the system and users requirements

Before Mat was designed, there were no techniques to measure or assess the quality of metadata in Greenstone. In order to enhance the functionality of Greenstone, we designed a system to help end-users maintain the quality of their collections. The first stage of software design is to define the goal of the project and draw requirements specification from end-users. The requirements define the functionalities or services that must be provided by the system. In general, the requirements should be complete and consistent. Completeness means that all services required by users should be defined; consistency means that requirements should not have contradictory definition. The list below presents the initial requirements specifications from end-users and repository managers.

- The system shall analyse an OAI repository and retrieve available metadata.
- The system shall build the Greenstone collections for users automatically.
- The system shall provide an overall statistical report and visualisation.
- The system shall allow users to view metadata records at different levels, such as repository-level and document-level.
- The system shall inform users when metadata records contain unusual values.

The basic architecture of the system can be derived from the requirements specification described above. The system is comprised from two key components: Greenstone and an analysis component. Greenstone is responsible for building collections and the analysis component is for analysing collections and generating statistical reports and visual displays. In section 3.2.3, a brief description of the development environment and detailed implementation rationale will be presented.

3.2.2 Use Cases

In order to verify that the program conforms to the requirements, it was decided to undertake a series of prototype based on four use cases derived from the requirements specification. The first was the implementation of a prototype as an infrastructure for a metadata quality tool. This was considered the most fundamental case of the four cases. The second case was to provide a visualisation of every metadata record; this system extends the functionality of Mat, by transforming the statistical data into a two-dimensional visual display. The third case was to provide a customised visualisation according to users' selections. The fourth case was to provide the online version of Mat. Due to the evolutionary nature of the development of the prototypes for each case, each implementation consisted of a similar architecture whereby harvested data was retrieved from a Greenstone digital library.

Use Case One

Use Case Name: Metadata Quality Tool System

Actor: repository manager

Design Goals: The number of digital resources has been increasing significantly in recent years. For a small digital repository, human experts can review collections manually. However, this approach is not applicable for large digital repositories due to numbers of records. In order to improve and maintain the quality of digital collections, repository managers are looking for an automatic metadata quality tool to assess their repositories. The main goal in this use case is to provide an infrastructure of the metadata quality tool. The system applies the related metadata quality metrics to measure the quality of metadata records and provides meaningful statistical data and visualisation for improving and maintaining the quality of digital repositories.

Preconditions:

The repository manager has used Greenstone to harvest OAI metadata records and build a new collection without any errors.

Basic course of events:

1. The repository manager executes the metadata quality tool and a message window will appear on the screen with a list of available collections.
2. The manager chooses one of the collections from the list and clicks “OK” button.
3. The system sends a message to Greenstone and receives information about that collection. Then the system starts to calculate quality metric values.
4. Once the statistical calculation for that collection has been done, a new window will appear on the screen with detailed information about that collection.

Alternative paths:

1. The repository manager clicks “Cancel” button to exit the system instead of choosing a collection.
2. The repository manager receives an explanation indicating why the operation cannot be done.

Use Case Two

Use case name: View Visualisation

Actor: repository manager

Design Goals: Repository managers may have ideas of how to improve quality for their repositories by reading statistical reports generated by Mat. However, the report does not indicate which records in the repository are empty or containing incorrect data. The report only contains basic statistics and lists of unique values. Therefore, repository managers hope to find a metadata quality tool that can provide visual displays of collections. The main goal in this use case is to provide the visualisations that allow users to analyse the collections.

Preconditions: The repository manager has used Greenstone to harvest OAI metadata records and build a new collection without any errors.

Flow of events:

1. The repository manager executes the program and selects a collection to analyse.
2. A new window will appear on the screen with statistical details of the collection.
3. The repository manager clicks “Visualisation” button to view the entire collection.
4. The system reads the file and transforms the values into a 2-dimensional visual display.
5. The repository manager magnifies an area of the visualisation by selecting the area.

Alternative paths:

1. If the repository manager does not click the "OK" button, the system will cancel the operation.
2. The repository manager receives an explanation indicating why the operation cannot be done.

Use Case Three

Use case name: View Customised Visualisation

Actor: repository manager

Design Goals: The overall visualisation provides every piece of metadata information to users. However, the visual interface requires lot of space to display the entire collection. Usually, users have to scroll the visual interface to find the metadata records they desire. The main goal in this scenario is to allow users to generate their own customised visualisation by selecting the metadata element they want to analyse.

Preconditions: The repository manager has used Greenstone to harvest the OAI metadata records and build the new collection for the records. The repository manager has executed the system and the main statistical report window appears on the screen.

Flow of events:

1. The repository manager switches to the individual metadata element set panel. The panel displays all metadata elements with their completeness value.
2. The repository manager selects the metadata elements they want to analyse

and clicks the “Visualisation” button.

3. The system shows a 2-dimensional visual display.
4. The repository manager can select a record from visualisation for a more detailed display.

Use Case Four

Use case name: Online Metadata Quality Tool

Actor: repository manager

Design Goals: Versions of Mat in the earlier cases require users to install Greenstone on their computers. It increases inconvenience for some people who do not have the permission to install software and decreases the usability of the metadata quality tool. The main goal in this use case is to generate an online version of the metadata quality tool for users to assess their repositories anytime and anywhere.

Preconditions: The repository manager is using a computer connecting to the Internet and the digital repository is available for metadata harvesting.

Flow of events:

1. The repository manager connects to the main page of Mat and provides an OAI URL to analyse.
2. The system sends a request to the OAI repository and retrieves the available metadata set.
3. The repository manager chooses one of the metadata set to analyse.
4. The system starts to download the metadata records and build the collection for further processing.

5. The repository manager is lead to a new web page with detailed information about the collection.

Alternative paths:

1. The OAI repository does not support OAI-PMH for harvesting. The system will display an error message to users with an explanation indicating OAI problems.

System Development

The Development Environment

The development environment consisted of a single personal computer with a single 2GHz CPU and 1 GB of RAM. The computer has installations of an Apache Tomcat server, a Greenstone3 digital library, and the GNU/Linux operating system. Greenstones3 uses Apache Tomcat web server to host its digital collections.

Implementation Rationale

The goal of this project was to find potential errors in metadata records in order to improve the quality of digital collections. In section 3.2.1, a basic structure of the system has been defined. Figure 20 shows the basic architecture of Mat (online version). This system is structured on top of Greenstone and uses a servlet (within Apache Tomcat server) to interact with users. The numbers in Figure 20 indicate the sequence of events in the system.

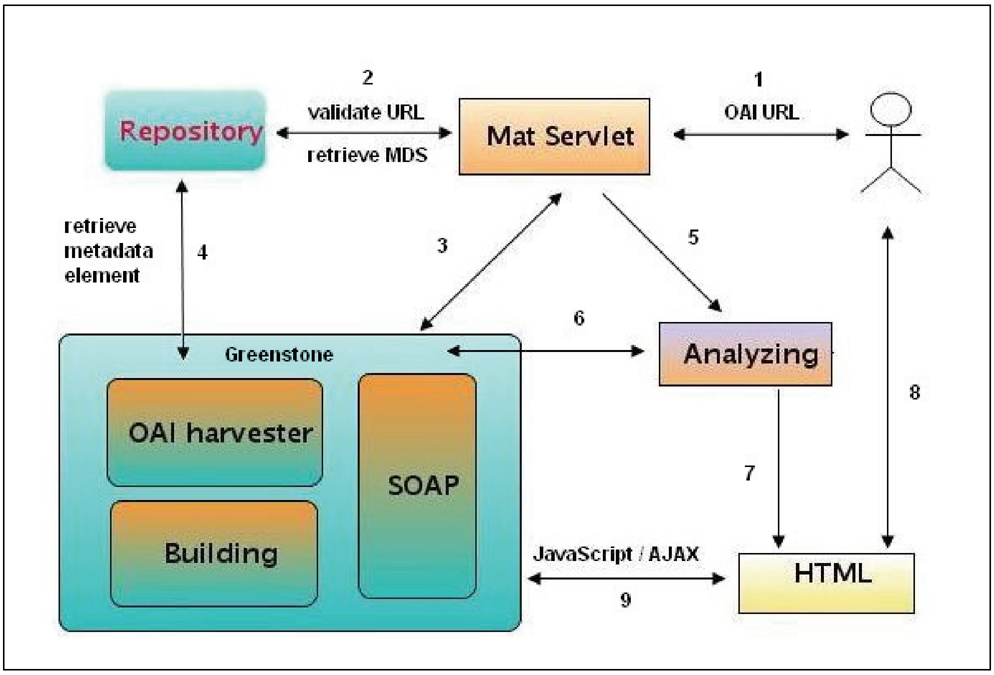
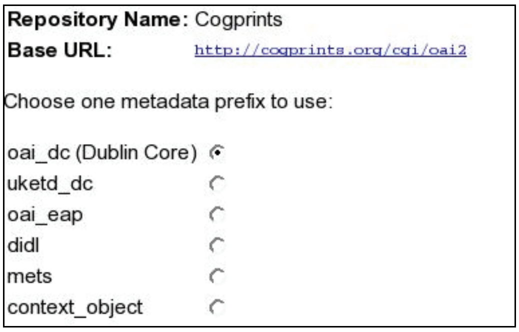


Figure 20 : Overview of the basic architecture

The design for the prototype metadata quality tool architecture uses the Mat servlet as a mediator between a data source (Greenstone) and the system core. Mat is made up of several units and consists of two key components:

- Greenstone digital library, used to download metadata records, build a collection and provide metadata information.
- An analysis component that implements the functionality of the metadata quality tool itself.

The servlet also plays a mediating role between Greenstone and the users. It validates URLs given by users and retrieves available collection-level metadata from digital repositories. Figure 21 is a screenshot of available metadata sets in Cogprints repository. This repository uses six types of metadata to index resources. Figure 22 is a part of an OAI response message from Cogprints (http://cogprints.org). The message lists all metadata sets that are used in Cogprints.

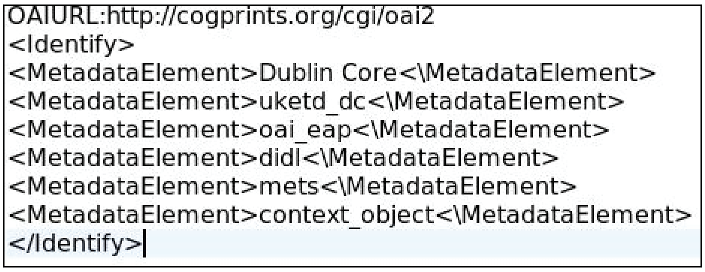


Repository Name: Cogprints
Base URL: http://cogprints.org/cgi/oai2

Choose one metadata prefix to use:

oai_dc (Dublin Core)	●
uketd_dc	○
oai_eap	○
didl	○
mets	○
context_object	○

Figure 21 : A screenshot of available metadata prefix of Cogprints



```
OAIURL:http://cogprints.org/cgi/oai2
<Identify>
<MetadataElement>Dublin Core<\MetadataElement>
<MetadataElement>uketd_dc<\MetadataElement>
<MetadataElement>oai_eap<\MetadataElement>
<MetadataElement>didl<\MetadataElement>
<MetadataElement>mets<\MetadataElement>
<MetadataElement>context_object<\MetadataElement>
</Identify>
```

Figure 22 : A screenshot of an OAI response message from the Cogprints repository

Then users can choose which metadata set they want to analyse. Greenstone is responsible for downloading OAI metadata records and building collections. Then the servlet calls the system core to analyse the collections. The system core retrieves information from Greenstone and uses it to calculate the quality metric values for each element. It is responsible for generating web pages and graphical user interface. Then the system core returns the URL of the report to the servlet and users can then access to the summary page of the collection report.

Figure 23 presents a flow chart of the prototype architecture of Mat. The system starts with downloading metadata records from the digital repository designated by users and then building the collection. Once the collection has been built successfully, the system core will be called. Then it starts to analyse the collections and store information into files. If any process cannot be completed, the system will stop the following operations and require users to return back to the main page.

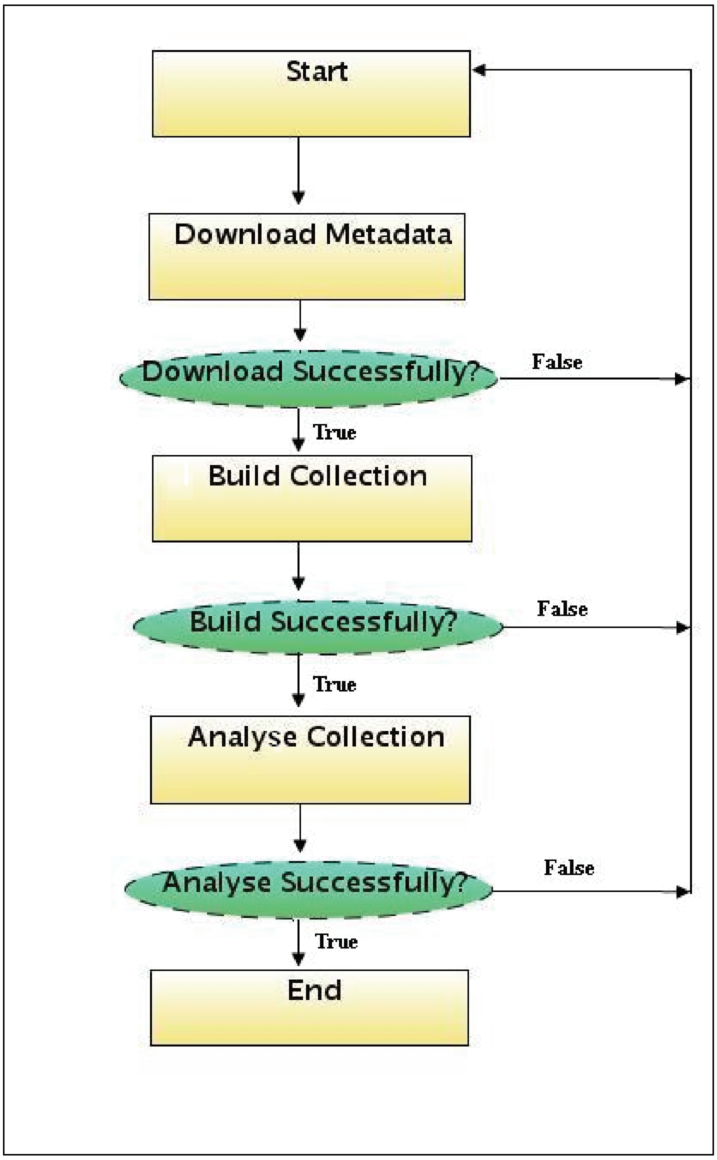


Figure 23 : Flow chat of Mat

Figure 24 shows the classes of the system, their inter-relationships, the operations and attributes. The analysis component is constructed from four main classes: the Data class, which calculates the statistics and stores the data; the Graphical User Interface class, which creates the GUI and handler users' request; the Visualisation class, which combines the statistics to draw the scatter plot/visualisation; the Printing class, which generates the WebPages.

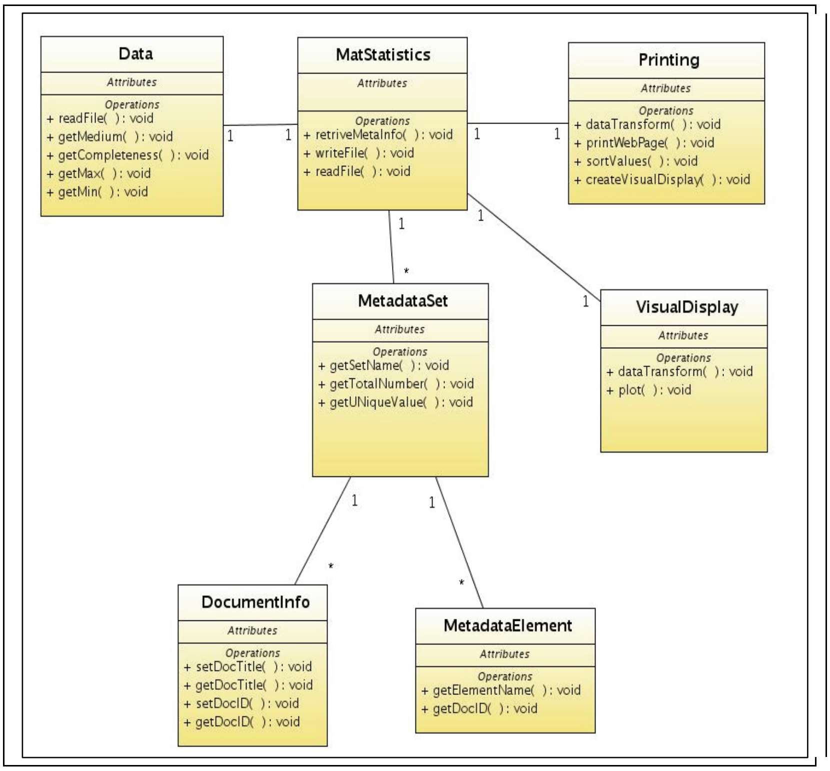
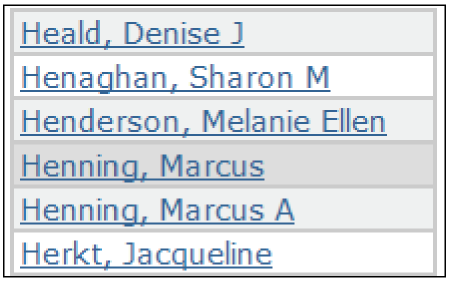


Figure 24 : Class diagram of the system

3.3 Duplicate Data Detection

Due to the large amount of data in metadata records, duplicate detection is an importance service. However, maintaining consistency is difficult. For digital libraries whose metadata is manually assigned by human experts, the issue of erroneous and duplicate metadata is particularly important. For large digital libraries, manual duplicate detection is infeasible and automated methods are necessary. To maintain quality digital libraries should constantly check their metadata.

Duplicate data detection is the technique of identifying multiple records that refer to the same object. The duplicate data detection focuses more on typographical errors and different representation of strings in the metadata rather than the records. Usually, the typographical variations of the string data are different by only one or two characters. For example, a user is using the online library to look for a book called "Using XML" but he enters "Using ZML". In the case, the server will return an empty list to the user. If the system has the ability to detect similar strings, users may find the book he/she desires. Our metadata tool is capable of calculating the similarity values for each unique string within the same group and generates a list of similar words for users to revise their metadata records. Using a controlled vocabulary can reduce the chance of typographical errors happening. Figure 25 is a screenshot of part of the author list of the AUT (Auckland University of Technology) digital repository. It is clear that there are two entries for the same author: "Henning, Marcus" and "Henning, Marcus A". This kind of error may confuse end-users when they are browsing the author list – and also makes the list longer. Therefore, we wish to develop a tool that could detect these data errors and report them to repository managers.



Heald, Denise J

Henaghan, Sharon M

Henderson, Melanie Ellen

Henning, Marcus

Henning, Marcus A

Herkt, Jacqueline

Figure 25 : Author list of AUT repository

Several well-known string similarity metrics have been developed to detect duplicates such as (Elmagarmid et al., 2007):

- Levenshtein distance
- Smith-Waterman distance
- Jaro distance
- Q-gram distance
- Affine gap distance

In this project, we use the Levenshtein distance to determine the similarity of strings. The Levenshtein distance is straightforward to implement and is suitable for the prototyping approach for exploring tool requirements.

3.3.1 The Levenshtein distance

The Levenshtein (Elmagarmid et al., 2007) distance is a character-based similarity metrics and relies on the string comparison technique to calculate the minimum number of edit operations needed to transform one string into the other. The idea is that for a misspelling, we should look for words that are relatively close. If the similarity of two strings exceeds the default threshold, the strings are considered different. Otherwise they are relatively close and may refer to the same thing. Therefore, it works well for detecting typographical errors.

The Levenshtein distance between two strings is the minimum number of edit operations of single character. The Levenshtein distance is an implementation of dynamic programming algorithm and permits three types of operation for transforming the source word to the target word.

- Insertion
- Deletion
- Substitution

For the Levenshtein distance, the cost of the deletions and insertions is one. The cost of substitution is one if the characters are different, otherwise it is zero. The following example illustrates how the Levenshtein distance works.

For example,

The words “computer” and “compute” are very similar and a change of just one letter, r->“_” will change the first word into the second (i.e. remove character “r”). The following table describes how the edit distance calculates the edit distance.

		C	**O**	**M**	**P**	**U**	**T**	**E**	**R**
	0	1	2	3	4	5	6	7	8
C	1	0	1	2	3	4	5	6	7
O	2	1	0	1	2	3	4	5	6
M	3	2	1	0	1	2	3	4	5
P	4	3	2	1	0	1	2	3	4
U	5	4	3	2	1	0	1	2	3
T	6	5	4	3	2	1	0	1	2
E	7	6	5	4	3	2	1	0	1

Table 4 : Explain how to calculate the Levenshtein distance

Step 1: We create a matrix with size of (length of string 1) * (length of string2).

In this case the size of matrix is 9 X 8.

Set n to be the length of string 1.

Set m to be the length of string 2.

Step 2: We initialise the first row to 0..n.

Step 3: We initialise the first column to 0..m.

Step 4: We examine each character of string1 (x from 1 to n).

Step 5: We examine each character of string2 (y from 1 to m).

Step 6: If s[x] equals t[y], then the cost is 0.

If s[x] doesn't equal t[y], then the cost is 1.

Ch1 means the current character of string1 and Ch2 means the current character of string2. It has to be explained in an edit of string1+ Ch1 into string 2+ Ch2.

- Ch1 equals Ch2, they are identical. No edit operation is required.
- Ch1 differs from Ch2, and then ch1 could be changed into ch2 by substituting one character. One edit operation required.
- Ch1 is not null but Ch2 is null, and then ch1 could be changed into ch2 by deleting a character. One edit operation required.
- Ch1 is null but Ch2 is not null, and then ch1 could be changed into ch2 by inserting a character. One edit operation required.

Step 7: The value of the matrix equal to the minimum of:

a. The cell immediately above plus 1.

b. The cell immediately to the left plus 1.

c. The cell diagonally above and to the left plus the cost.

The running time for Levenshtein distance algorithm of Table 4 is O(mn): n is the length of string1, m is the length of string2. However, this simple implementation of the algorithm can be improved to run in O(m) by discarding earlier rows. Figure 26 and Figure 27 show this improvement. The vertical axis represents the time in milliseconds while the horizontal axis represents the number of records.

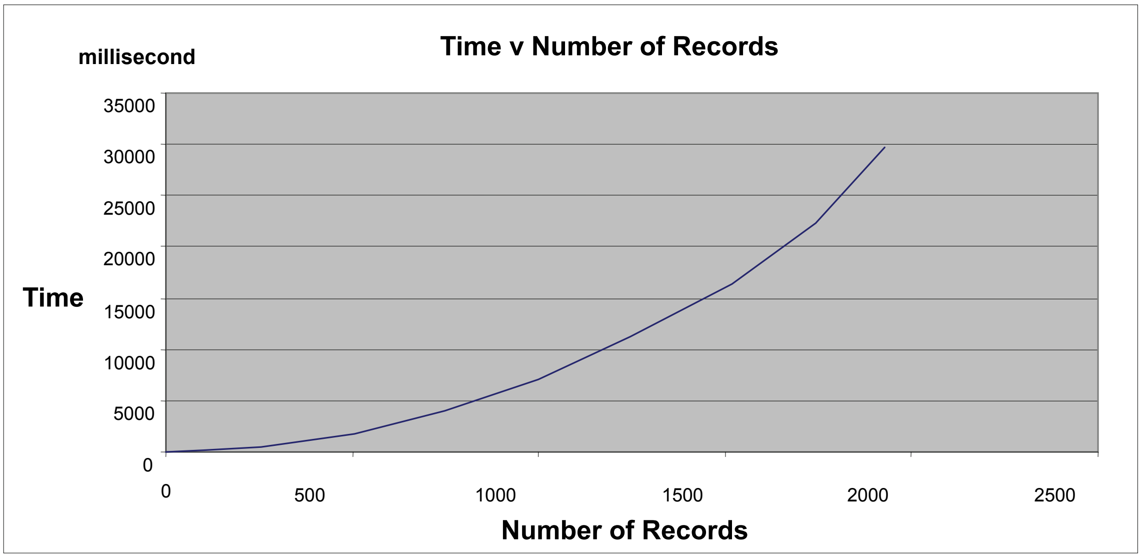


Figure 26 : Time taken for using the original edit distance algorithm.

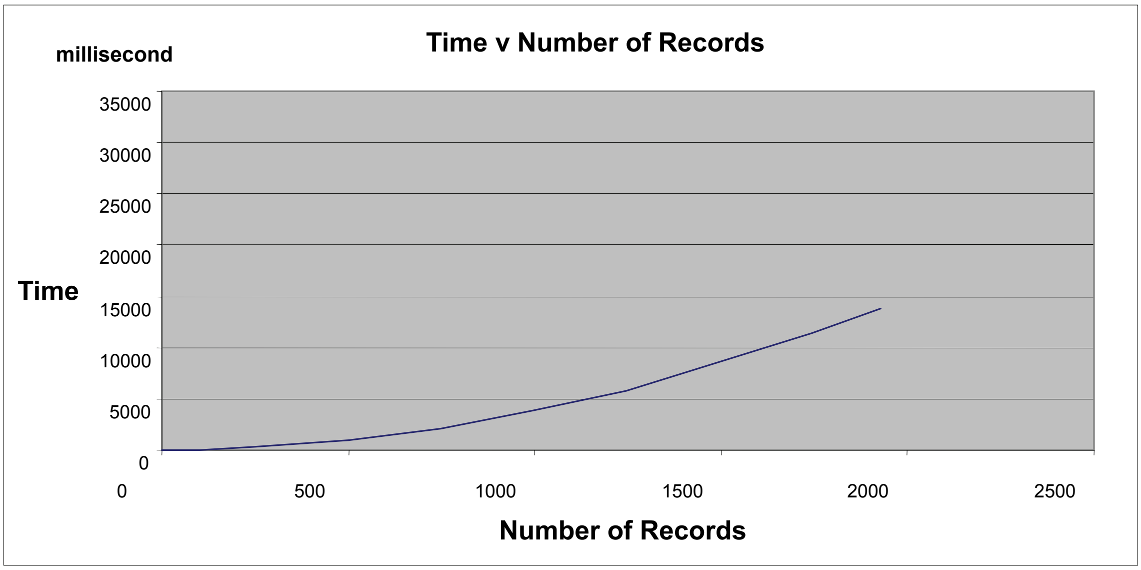
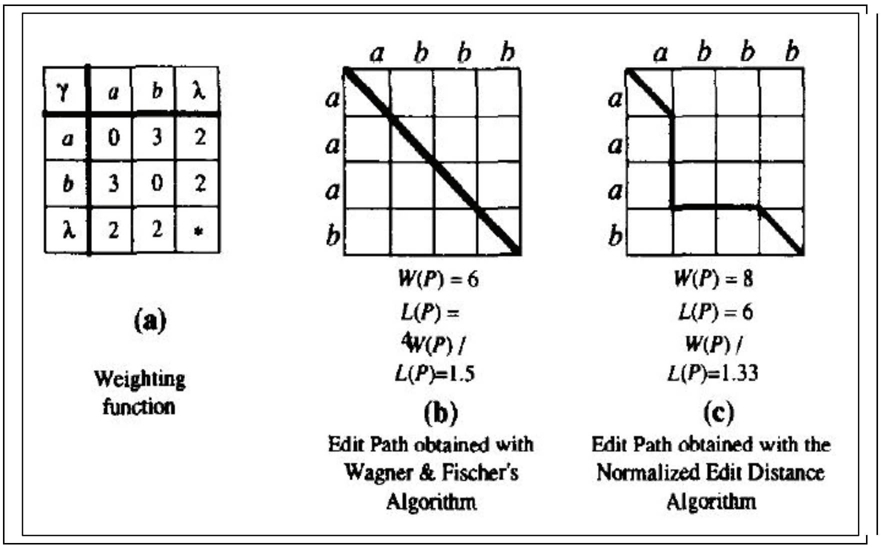


Figure 27 : Time taken for using the modified edit distance algorithm.

"However, the Levenshtein distance is not totally suitable for every applications since it lacks some type of normalization that would appropriately rate the weight of the (edit) errors with respect to the size of the objects (string) that are compared" (Marzal and Vidal, 1993).

For instance, two (edit) errors in the comparison between strings of length 3 are more important than three errors in a comparison of strings of length 6.

> *"The normalized edit distance between X and Y is defined as the minimum number of edit W(P)/L(P), where P is an editing path between X and Y, W(P) is the sum of the weights elementary edit operations of P. L(P) is the number of these operations(length of P). It is defined directly in terms of paths rather than transformations. In fact, unless certain nontrivial conditions are imposed on the elementary edit weight function r and/or on the definition of edit sequences, no meaning definition of normalized edit distance seems possible in terms of edit transformations" Marzal and Vidal, 1993).*



(a) Weighting function

(b) Edit Path obtained with Wagner & Fischer's Algorithm

(c) Edit Path obtained with the Normalized Edit Distance Algorithm

Figure 28 : Example of edit distance with post-normalization versus normalized edit distance from Marzal and Vidal (1993)

Figure 28 shows the difference between the post-normalization and normalized edit distance. Because the difference between the unnormalized edit distance and normalized edit distance is very close, the normalized edit distance does not seem to give a large improvement in detecting duplicates. Therefore, the normalized edit distance is not used at this stage of the study.

3.3.2 Phonetic Similarity Metrics

The Levenshtein distance focuses on the string representations of the metadata records. However, string may be pronounced identically even if their representations are very different. Even common names can be misspelled due to minor variations in spelling, for instance “Smith” and “Smyth” are two variations on a common name that sound the same. The phonetic similarity metrics are trying to address such issues and match strings. There are several phonetic metrics have been invented such as Soundex, New York State Identification and Intelligence System (NYSIIS), Oxford Name Compression Algorithm (ONCA), Metaphone and Double Metaphone (Elmagarmid et al., 2007; White 2004; Navarro 2001).

Soundex Algorithm

“Soundex is the most common phonetic similarity metrics and has been used to index all US censuses from 1920 onwards” (White, 2004). It was invented by Margaret K. Odell and Robert C. Rusell in 1918. The algorithm tends to group the names phonetically rather than according to the alphabetic construction of the names. It maps each letter to a numerical code representing its phonetic group.

Letter	Phonetic group.
B, F, P, and V	1
C, G, J, K, Q, S, X, and Z	2
D and T	3
L	4
M and N	5
R	6
A, E, I , O, U, and Y	0

Table 5 : Mapping between letter and phonetic group

“Newcombe reports that the Soundex code remains unchanged, exposing about two-third of the spelling variations observed in the linked pairs of the vital records” (Elmagarmid et. al, 2007). The phonetic similarity metrics are suitable for detecting misspelling mistakes in personal names. As with more advanced string similarity measures we leave phonetic approach for future work.

3.4 Summary

This chapter has explained the means by which our Mat architecture was implemented. Section 3.2.1 provided a brief introduction of the system and requirements specification from end-users and repository managers. Section 3.2.2 presented a series of use cases to give the guidance in the development of the system. Section 3.2.3 presented overviews of the development environment and detailed implementation rationale of the system. Finally, Section 3.3 discussed the deduplication techniques and how to adapt algorithm to fit our system. The following chapter will discuss the evolution of Mat.

Chapter 4

Mat Evolution

4.1 Introduction

The first version of Mat was deployed in January 2008. Since then it has been upgraded to a second version with several improvements. These improvements were based on initial feedback and suggestions from repository managers. We used an online survey and semi-structured interviews with repository managers to obtain first-hand feedback. In this chapter we will introduce some key features of the initial and current versions.

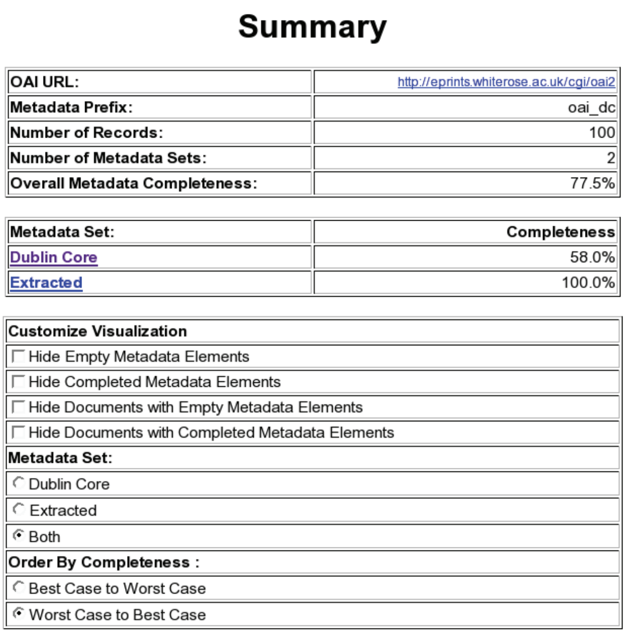
4.2 Mat Alpha Version One

We identified a number of requirements as we developed our first prototype of Mat in October 2007 and developed a system that would comply with the user requirements. Mat was designed to assist both end-users and repository managers to improve and maintain metadata quality of their collections. This tool generated detailed statistical reports of collections and completeness-oriented visual displays similar to "Spotfire" (see section 2.5.3) to demonstrate the overall distribution of metadata in a collection. The statistical reports provided usage of metadata sets and elements to help repository managers maintain completeness. The visual displays allowed repository managers to examine metadata elements and find unusual metadata records. Appendix B shows screenshots of both the early Java prototype and the first online version of Mat.

This tool utilised some well-known metadata quality metrics to compute statistics for collections, metadata sets and metadata elements and generated reports to express these values. The statistical reports comprise three main sections:

- summary description of entire collections
- usage of metadata sets and elements
- sorted lists of unique metadata values

The summary description provided basic information about collections and lists all available metadata sets. It usually comprised two metadata sets: Extracted Metadata Set and Simple Dublin Core Metadata Set. This extracted metadata set is generated automatically when documents are imported into Greenstone and the Dublin Core Metadata Set is used by the repositories. Qualified Dublin Core metadata is also supported by this system. Figure 29 shows a summary page of a report. The first section shows basic information of an OAI repository and metadata records. The second section indicates the metadata set used in this collection. The last section lists all available options for repository managers to create customised visualisations. Clicking the “Dublin Core” link takes users to Figure 30. Figure 30 shows 15 Dublin Core elements and six of them are complete (100 % completeness).



Summary

OAI URL:	http://eprints.whiterose.ac.uk/cgi/oai2
Metadata Prefix:	oai_dc
Number of Records:	100
Number of Metadata Sets:	2
Overall Metadata Completeness:	77.5%

Metadata Set:	Completeness
Dublin Core	58.0%
Extracted	100.0%

Customize Visualization

□ Hide Empty Metadata Elements

□ Hide Completed Metadata Elements

□ Hide Documents with Empty Metadata Elements

□ Hide Documents with Completed Metadata Elements

Metadata Set:

○ Dublin Core

○ Extracted

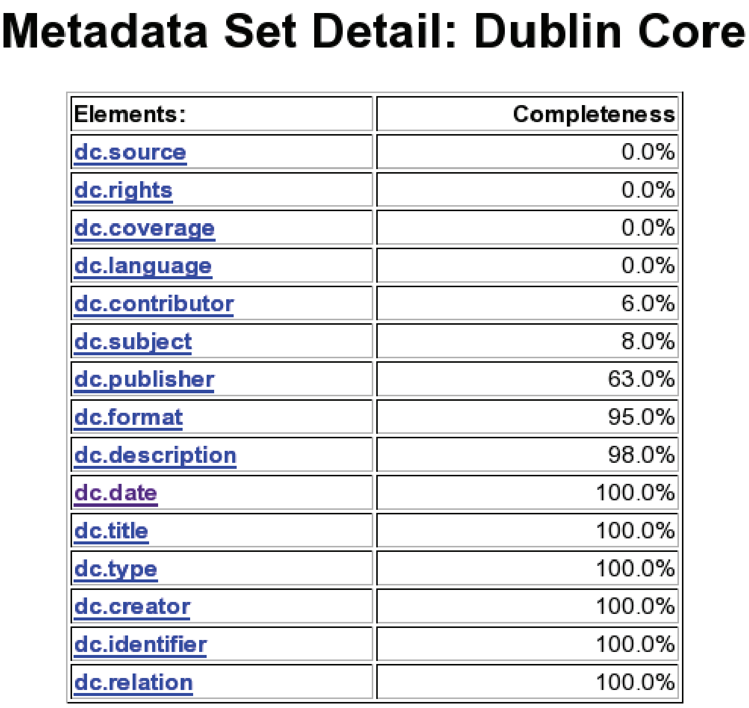
● Both

Order By Completeness :

○ Best Case to Worst Case

● Worst Case to Best Case

Figure 29 : The summary description of 100 OAI metadata items

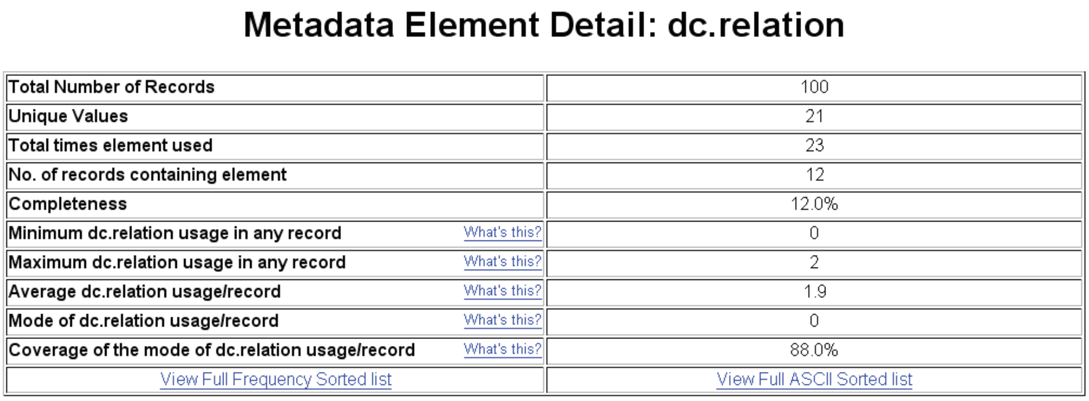


Metadata Set Detail: Dublin Core

Elements:	Completeness
dc.source	0.0%
dc.rights	0.0%
dc.coverage	0.0%
dc.language	0.0%
dc.contributor	6.0%
dc.subject	8.0%
dc.publisher	63.0%
dc.format	95.0%
dc.description	98.0%
dc.date	100.0%
dc.title	100.0%
dc.type	100.0%
dc.creator	100.0%
dc.identifier	100.0%
dc.relation	100.0%

Figure 30 : A metadata view of 15 Dublin Core elements

Each metadata element page has some descriptive statistical measures and two different types of sorting methods: ASCII and frequency sorting. These quality values are calculated by using the formula provided by Duval and Ochoa (2006). Figure 31 shows summary statistics of the “dc.relation” element. This element has 21 different metadata values and average usage per record is about 1.9 (23 times / 12 records).

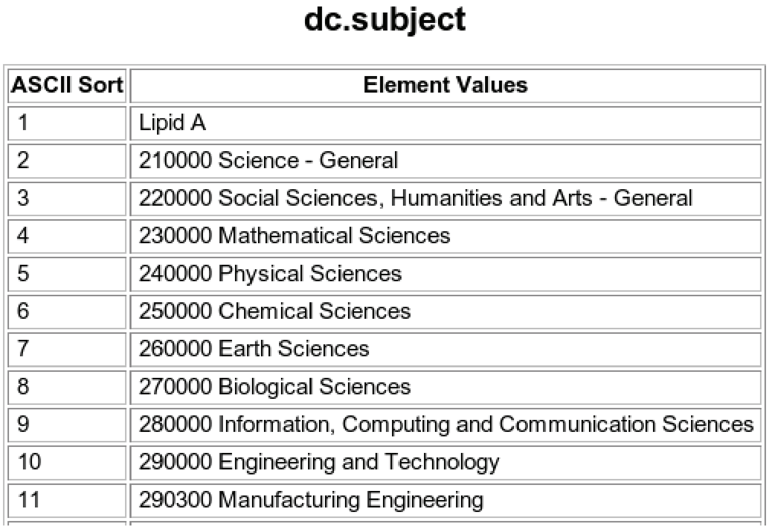


Metadata Element Detail: dc.relation

Total Number of Records		100
Unique Values		21
Total times element used		23
No. of records containing element		12
Completeness		12.0%
Minimum dc.relation usage in any record	What's this?	0
Maximum dc.relation usage in any record	What's this?	2
Average dc.relation usage/record	What's this?	1.9
Mode of dc.relation usage/record	What's this?	0
Coverage of the mode of dc.relation usage/record	What's this?	88.0%
View Full Frequency Sorted list		View Full ASCII Sorted list

Figure 31 : Part of the element detail view

The sorting techniques are used to list metadata values and allow users to view every single unique value at the same time. There are two types of sorting technique offered by statistical reports: ASCII sorting and frequency sorting. The ASCII sorting is designed to allow terms that begin with unusual characters to “float to the top” or “sink to the bottom” of the list. The frequency sorting is for finding the most popular terms used in element metadata values. Figure 32 displays a part of the ASCII sorting list of a “dc.subject” element.

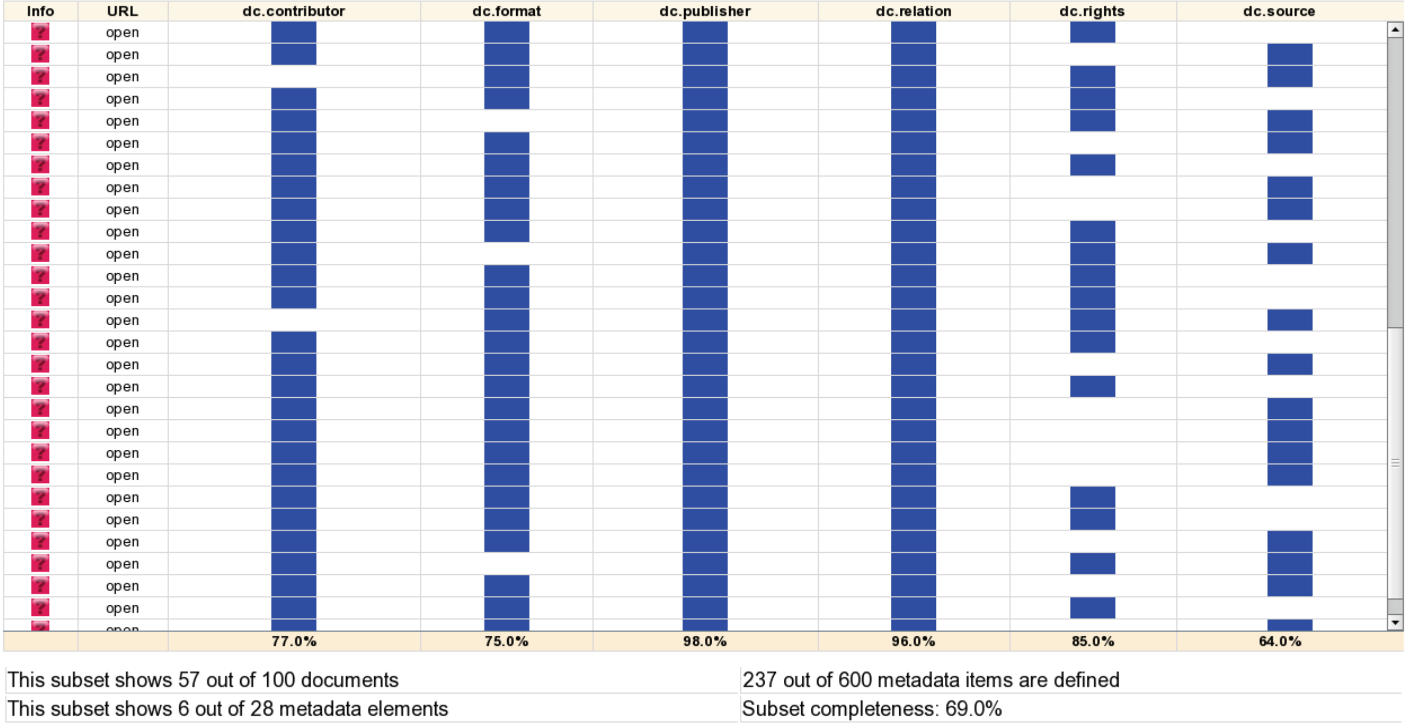


dc.subject

ASCII Sort	Element Values
1	Lipid A
2	210000 Science - General
3	220000 Social Sciences, Humanities and Arts - General
4	230000 Mathematical Sciences
5	240000 Physical Sciences
6	250000 Chemical Sciences
7	260000 Earth Sciences
8	270000 Biological Sciences
9	280000 Information, Computing and Communication Sciences
10	290000 Engineering and Technology
11	290300 Manufacturing Engineering

Figure 32 : A part of an ASCII sorting list

The visualisation displays of our system provide simple scatter plots of overall distribution of collections. The horizontal axis represents metadata elements while the vertical axis represents metadata items. Figure 33 shows six Dublin Core metadata elements (as eight completed elements have been hidden) and 100 metadata items in the table. The intersection of the X and Y axis indicates whether this metadata element is defined for this metadata item. If this metadata element is defined, a blue rectangle is used to indicate the present of that metadata item. Otherwise, a white rectangle will be used to indicate this metadata element is not defined for this metadata item. The column on the left of Figure 33 showing question marks are used to display the full metadata record for resources. The "URL" column contains a list of links back to resources in the remote repository.



This subset shows 57 out of 100 documents

This subset shows 6 out of 28 metadata elements

237 out of 600 metadata items are defined

Subset completeness: 69.0%

Figure 33 : A scatter plot of 100 records and 6 Dublin Core metadata elements

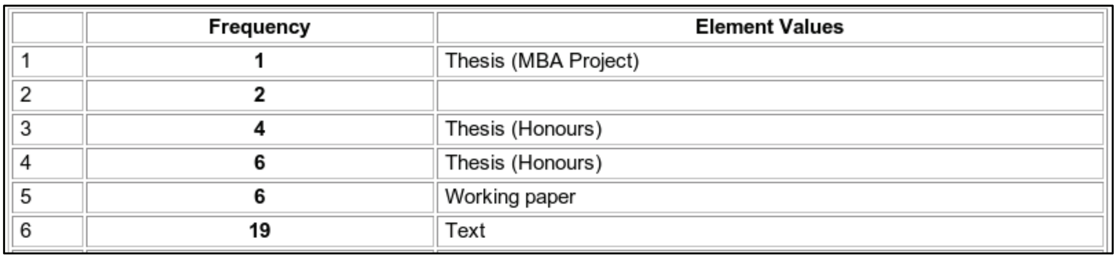
Findings

During the development of Mat, we discovered an unusual problem of Greenstone's metadata extraction techniques. In general, Greenstone would try to extract file content and create an extracted metadata record for each object. Every document/record should have an extracted metadata value for its title. However, we discovered that not every document/record had an extracted title value in some situations. One report showed that extracted title elements did not have 100 percent completeness. The tool's reports alerted the Greenstone development team to a specific case where titles were not assigned.

A second finding was Greenstone did not close its database properly in some circumstances. Our system was working initially, but then we received "too many open files" errors from the Apache Tomcat server. This error could cause the server to shut down and decrease stability of our system. Every time Mat accessed the Greenstone database, it opened every collection but did not correctly close the associated resources. If the number of open files is more than a system limit, the system would shut down and no service would be available until the server is restarted. The Greenstone development team located and fixed the problem and the stability of our system was improved significantly.

The third finding was that several reports showed that some repository managers misused the Dublin Core metadata elements. For example, repository administrators put "text" in the "dc.type" elements. The "dc.format' describes the file format and physical medium; "dc.type" defines the genre of the resource. In this situation, repository managers should use "dc.format" to describe the genre of the records instead of "dc.type". We found a large degree of inconsistency in the use of Dublin Core metadata elements.

The fourth finding was that a number of metadata records contained empty/white spaces. As seen in Figure 34, the dc.type element comprises of six unique values and five percent of total values are empty/white spaces. As a result of using HTML to present unique value lists, the list looks unusual.



	Frequency	Element Values
1	**1**	Thesis (MBA Project)
2	**2**	
3	**4**	Thesis (Honours)
4	**6**	Thesis (Honours)
5	**6**	Working paper
6	**19**	Text

Figure 34 : A part of a frequency sorted list of dc.type element

Finally, despite these issues the tool proved useful to repository managers:

> *"During one interview a participant noticed that an element in her repository had a non-zero completeness value when the local policy was not to use the element at all. Although the current tool doesn't provide a link to the affected records, she simply copied the value, searched the repository, located the records and then corrected them using the web administration interface of the repository" (Nichols and Chan et al., 2008).*

4.3 Mat Alpha Version Two

After our prototype was completed we began to improve the functionality of the tool. As mentioned earlier, we used an online survey and interviews with repository managers to gather feedback. Many of these improvements are based on user feedback. Mat alpha version two includes six major improvements over the previous version, including fixes for stability and system response time. In this section, we will introduce these improvements and discuss related findings.

1. Remove Greenstone Extracted Metadata Set

In the previous version, the tool always reported two metadata sets in its results. This extracted metadata set is not that useful and valuable to end-users and repository managers and its completeness is usually 100 percent. The system provided four types of filtering methods, three types of metadata sets (Extracted, Dublin Core, and Both), and two types of ordering approaches. It needed to generate 96 possible visual displays in advance and this process took considerable time. Therefore, some repository managers suggested that the system should remove the extracted metadata set to reduce the amount of workload and system response time.

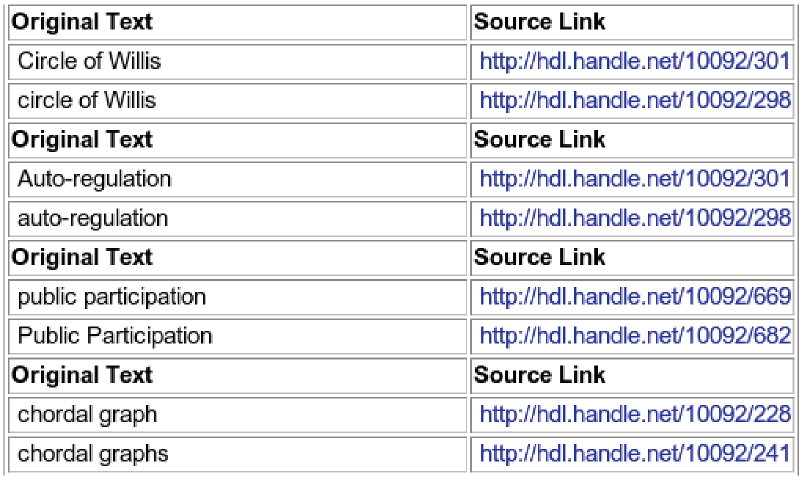
In the current version, we removed this unvalued extracted metadata to save system processing time. Now, the system offers one metadata set to users and only needs to generate 32 possible visual displays. Average user waiting time is significantly reduced.

2. Improve ASCII Sorted List

In section 4.2, we saw that if identifier URLs were provided to repository managers, they would not need to copy metadata values and search the repository. As shown in Figure 32, the original ASCII list did not have the source links and internal metadata displays. This idea was implemented in the second version. Users can now use source links back to original documents in remote OAI repositories and examine metadata elements. This feature not only assists repository managers to find documents more easily but also helps developers to detect errors. Figure 38 is a part of improved ASCII sorted list in alpha version two.

3. Potential Duplicate List

One of the most important features of alpha version two is the potential duplicate list (using the technique described in section 3.3.1). It provides a list of similar words that may be the same but the string representations are not identical. The idea of this functionality is to distinguish every metadata value and find any possible connections between them. Two words would be considered different if their string representations are not identical. Figure 35 shows a small part of a potential duplicate list. The list shows various types of different string representations for metadata values. For example, “Circle of Willis” and “circle of Willis” should be the same. However, the second metadata value does not start with a capitalized letter C. This list could assist repository managers to create a controlled vocabulary list and ensure that each concept is described using only one authorised term.



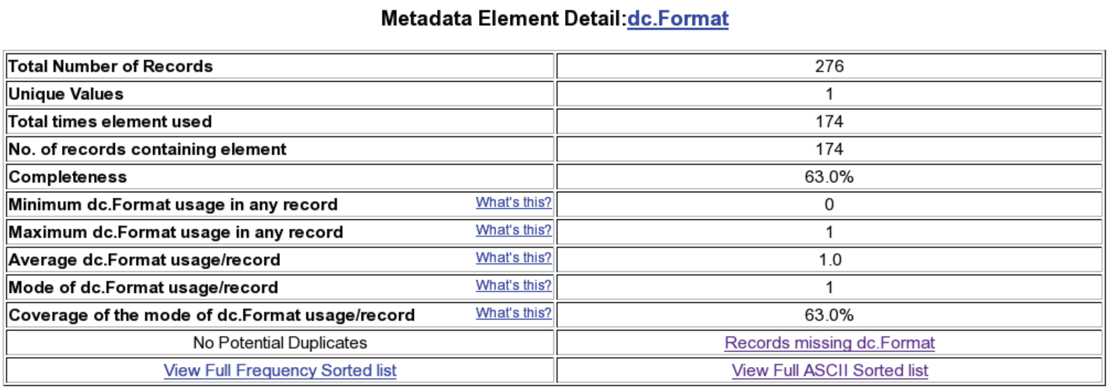
Original Text	Source Link
Circle of Willis	http://hdl.handle.net/10092/301
circle of Willis	http://hdl.handle.net/10092/298
Original Text	**Source Link**
Auto-regulation	http://hdl.handle.net/10092/301
auto-regulation	http://hdl.handle.net/10092/298
Original Text	**Source Link**
public participation	http://hdl.handle.net/10092/669
Public Participation	http://hdl.handle.net/10092/682
Original Text	**Source Link**
chordal graph	http://hdl.handle.net/10092/228
chordal graphs	http://hdl.handle.net/10092/241

Figure 35 : A part of a potential duplicate list

4. Links to Dublin Core Element and New Missing List

As mentioned in section 4.1, some repository managers misused the Dublin Core metadata elements. If that metadata element plays an important role in browsing or searching, the consequences of misinterpreted or misused metadata element are potentially large. One repository manager suggested links to the Dublin Core Metadata Initiative descriptions of metadata types in the element detail page would be helpful.

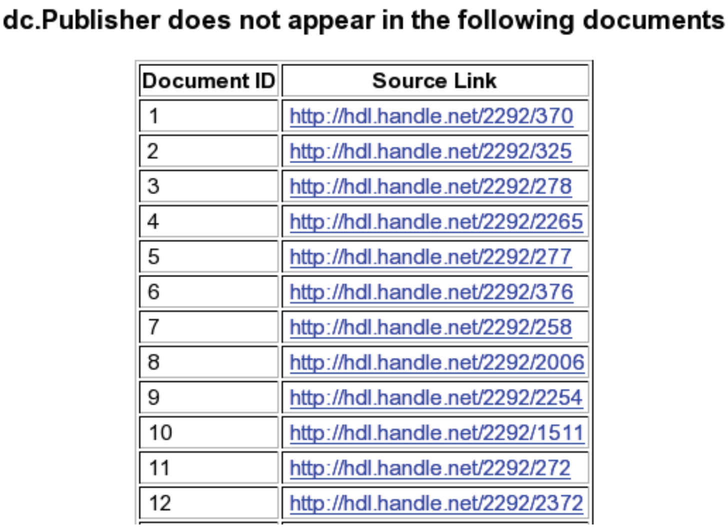
Figure 37 shows a list of records which do not contain the “dc.publisher” elements. In the previous version, statistical reports only showed an overall completeness of that collection and users would have to use visual displays or other tools to find those incomplete metadata items. The missing list is another new feature of alpha version two. A missing list of metadata element indicates which record does not define that element. Figure 36 shows that 37% of 276 items do not define the “dc.Format” elements. If repository managers want to find those incomplete metadata items, previously they would have to use their administration tools. This new feature allows repository managers to locate these items easily and efficiently.



Metadata Element Detail:dc.Format

Total Number of Records		276
Unique Values		1
Total times element used		174
No. of records containing element		174
Completeness		63.0%
Minimum dc.Format usage in any record	What's this?	0
Maximum dc.Format usage in any record	What's this?	1
Average dc.Format usage/record	What's this?	1.0
Mode of dc.Format usage/record	What's this?	1
Coverage of the mode of dc.Format usage/record	What's this?	63.0%
No Potential Duplicates		Records missing dc.Format
View Full Frequency Sorted list		View Full ASCII Sorted list

Figure 36 : A screenshot of the metadata element detail page



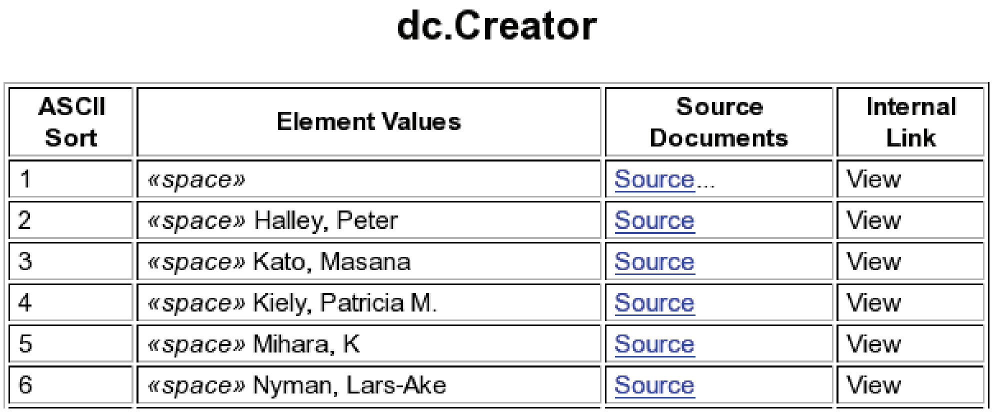
dc.Publisher does not appear in the following documents

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
4	http://hdl.handle.net/2292/2265
5	http://hdl.handle.net/2292/277
6	http://hdl.handle.net/2292/376
7	http://hdl.handle.net/2292/258
8	http://hdl.handle.net/2292/2006
9	http://hdl.handle.net/2292/2254
10	http://hdl.handle.net/2292/1511
11	http://hdl.handle.net/2292/272
12	http://hdl.handle.net/2292/2372

Figure 37 : A part of a missing list

5. Using Special Symbols to Indicate Problems

In the previous version of Mat, the content of metadata records was represented in its original format. However, this may cause ambiguity for some users whose metadata records contain whitespace. An example of this problem can be seen in Figure 38. Previously, there was no indication to alert users to whitespace problems. Therefore, the current version of Mat has fixed this problem by adding special symbols to notify users. The “<<*space*>>” symbol is used to indicate that the element values contain white spaces.



dc.Creator

ASCII Sort	Element Values	Source Documents	Internal Link
1	*«space»*	Source...	View
2	*«space»* Halley, Peter	Source	View
3	*«space»* Kato, Masana	Source	View
4	*«space»* Kiely, Patricia M.	Source	View
5	*«space»* Mihara, K	Source	View
6	*«space»* Nyman, Lars-Ake	Source	View

Figure 38 : An improved ASCII sorted list

Findings

Invalid OAI XML causes the Greenstone Building Process to Fail

This problem was discovered when we tried to analyse one particular institutional repository. The problem was that Greenstone harvested invalid OAI XML files from the repository but did not stop its collection building process. Hence, Mat servlet received signals from Greenstone and the system started to analyse the collection. Some users reported they could not find a Dublin Core link on the summary pages. Because the collection was not built properly, the system would not be able to analyse it. In order to notify users the repository is sending back invalid OAI XML files, we noted that Mat servlet should provide an explanation of this problem.

4.4 Summary

This chapter has introduced some basic features of previous and current versions. In section 4.2, the descriptions and screenshot of the statistical report and visualisation were given. The statistical reports had three main features:

- summary description of collections
- usage of metadata sets and elements
- sorted lists of unique values.

The visual displays provided scatter plots of overall distribution and allowed users to explore metadata elements and items. We used an online survey and interviews to gather user feedback after the tool was released to the public in January 2008. Section 4.3 discussed some important improvements in alpha version two. The most important improvements are the potential duplicate list and ASCII list. These two features are specially designed for finding errors in metadata element values.

Chapter 5

Discussion

5.1 Introduction

The Mat tool has generated more than 300 reports for institutional repositories since the deployment in January 2008. In section 5.2, we will present several common metadata errors/mistakes. As mentioned in chapter 2, most institutional repositories use the simple Dublin Core to index their collections. “Due to the flexibility of simple Dublin Core, there is considerable variation in how repository managers use these metadata elements” (Jordan, 2006). We will analyse the use of the Dublin Core Metadata Element Set in institutional repositories in section 5.3. Qualitative feedbacks from repository managers will be presented in section 5.4.

5.2 Potential Duplicates

In this section we discuss the common mistakes/errors found in metadata records. The mistakes/errors can be categorized into four types by analysing 22,200 metadata records from 20 different institutional repositories. In the following sections, we will present the common errors and discuss how Mat handles these errors. In order to compare the data generated by Mat with the real metadata, each sub-section has two figures. The first figure is a screenshot of the potential duplicates list generated by Mat and the second figure is the view from the remote repository.

5.2.1 Spacing Errors

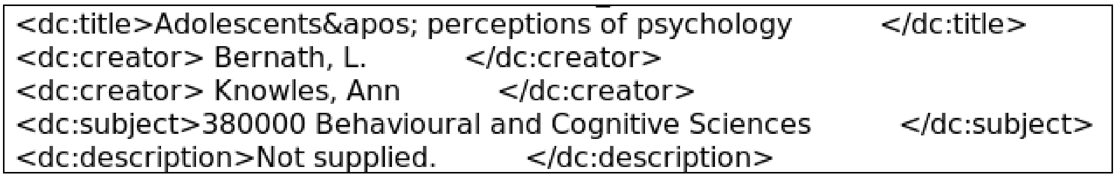
We found the “spacing” problem is one of the most common errors in metadata records. Repository managers usually use white spaces to separate author’s first name and middle name. However, there is no rule defining how many spaces should be placed between them. Most repository managers place one space to separate words but some repository managers do not use it. Therefore, the consistency and quality of digital repositories is hard to maintain. These issues derived from the lack of authority control in repository software (Nichols and Chan et al., 2008). Several types of mistakes about the “spacing” will be introduced in this section. Firstly, we will discuss problems of leading and trailing spaces and followed by solutions to this problem. Secondly, we will discuss the spaces between words.

Leading and Trailing Spaces

As shown in Figure 39, there is a white space character at the beginning of the “dc.creator” element. This issue may be from human data entry or from transforming metadata values to different metadata sets. This mistake may not be found easily on web pages because HTML treats a sequence of white-space characters as a single space.

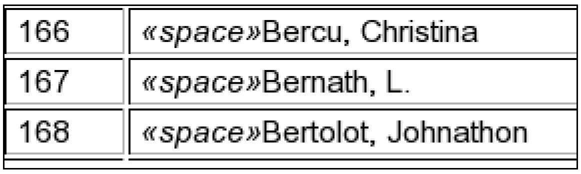
In general, it is useless to compare a regular string leading and trailing space characters. For example, there are two strings: string A is “Hello World” and string B is “   Hello World” with 4 leading space characters. Obviously, string A and String B are the same but the edit distance between string A and string B is four. In this situation, string A and string B are considered different. Due to this leading and trailing space problem, the simple edit distance approach of Mat was unable to provide correct potential duplicate lists.

If a string contains leading space characters, they will be removed with some extra edit distance costs. The default edit distance cost for a white-space character is 0.2. After the edit distance has been calculated, Mat needs to restore the sting to its original form and generates the potential duplicate list. As mentioned earlier, any sequence of white-space characters is treated as a single space. Therefore, users cannot distinguish string A and string B on the web page. A special indicator "<<*space*>>" is used to help users notice this spacing problem. Figure 40 shows an example of this problem. The white-space character is replaced by a special indicator. Any leading and trailing space character is replaced by "<<*space*>>".



```
<dc:title>Adolescents&apos; perceptions of psychology        </dc:title>
<dc:creator> Bernath, L.          </dc:creator>
<dc:creator> Knowles, Ann          </dc:creator>
<dc:subject>380000 Behavioural and Cognitive Sciences        </dc:subject>
<dc:description>Not supplied.          </dc:description>
```

Figure 39 : A part of an OAI record



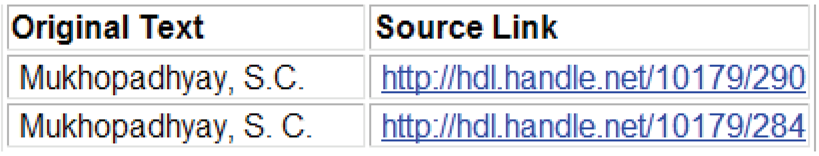
166	«*space*»Bercu, Christina
167	«*space*»Bernath, L.
168	«*space*»Bertolot, Johnathon

Figure 40 : Screenshot of a potential duplicate list – Metadata values containing leading spaces

Spaces between Words

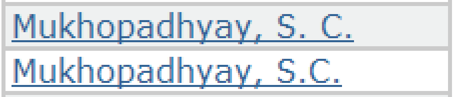
As shown in Figure 41, there are no spaces to separate the author's name in the first row. When the repository manager entered “Mukhopadhyay S.C.” to the digital repository, he/she did not use a white space character to separate S and C. However, this name was entered to the repository again with a space character between these two letters. Although, the string representations of these two names are different, they refer to the same author. As a consequence, Mukhopadhyay’s papers are divided and mislabelled into two different duplicate author entries (See Figure 42).

In this example, there is only one white space character between these two letters. If there are multiple spaces between words, then the pre-processing will try to merge spaces into one single space with some edit distance costs. Only once this is done will Mat start to compare the strings.



Original Text	Source Link
Mukhopadhyay, S.C.	http://hdl.handle.net/10179/290
Mukhopadhyay, S. C.	http://hdl.handle.net/10179/284

Figure 41 : A screenshot of a potential duplicate list



Mukhopadhyay, S. C.
Mukhopadhyay, S.C.

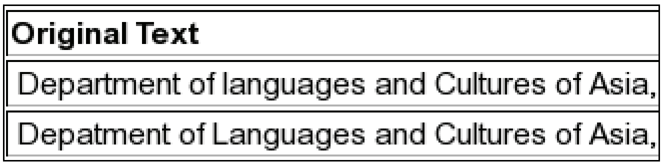
Figure 42 : A screenshot of an author list – two entries referring to the same person (spacing problem)

5.2.2 Typographical Errors

Incorrect data come in various forms and a typographical error is one of them. A typographical error is a mistake made during the typing process by pressing a wrong key on a keyboard. As shown in Figure 43, the word “department” is spelled in two different ways and there is a character missing in the second one.

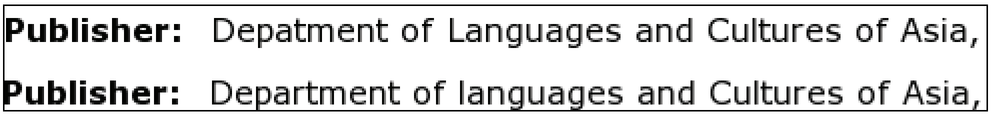
Typically, the typographical error is more serious than other metadata errors. “The accuracy is the degree to which the metadata elements match the objects” (Duval and Ochoa, 2006). The leading space character problem mainly affects the representation of metadata values. However, typographical errors always affect metadata values. Hence, we focus more on detecting typographical errors than other metadata mistakes.

As mentioned in section 3.3.1, Mat uses the Levenshtein distance (Elmagarmid et al., 2007) to catch most metadata errors. Mat could easily find typographical errors by applying the Levenshtein distance. However, there is a limitation; if the word “Department” does not exist in any other metadata records, Mat will not be able find this error.



Original Text
Department of languages and Cultures of Asia,
Depatment of Languages and Cultures of Asia,

Figure 43 : A screenshot of a potential list with a spelling error



Publisher: Depatment of Languages and Cultures of Asia,

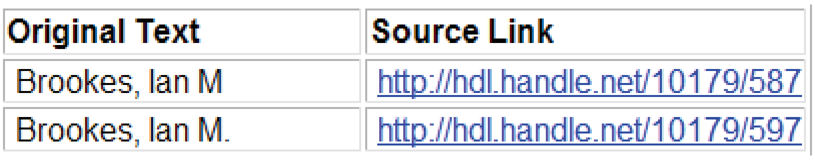
Publisher: Department of languages and Cultures of Asia,

Figure 44 : A screenshot of the “Publisher” elements – Department is misspelled as “Depatmenet” (typographical error)

5.2.3 Punctuation Errors

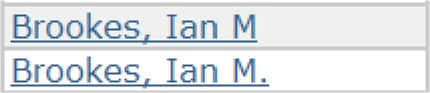
The punctuation mistake is similar to the space mistake and may be caused by accident. For peoples' names, many people tend to use the abbreviation for their middle names. "IDRC's (International Development Research Centre, 2008) style is to use few periods (full stops) in abbreviations". Figure 45 shows an example of this metadata error. The name in the first row does not have the full stop but the name in the second row uses the full stop.

In this situation, one edit distance operation is required for transforming name in first row into name in the second row. The punctuation does not affect the accuracy of the metadata value very much.



Original Text	Source Link
Brookes, Ian M	http://hdl.handle.net/10179/587
Brookes, Ian M.	http://hdl.handle.net/10179/597

Figure 45 : A screenshot of a potential list



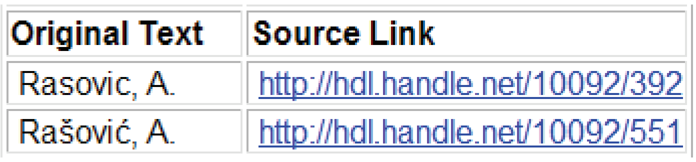
Brookes, Ian M

Brookes, Ian M.

Figure 46 : A screenshot of an author list – two entries referring to the same person (punctuation error)

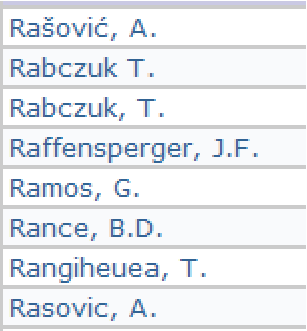
5.2.4 Diacritic Errors

A diacritic is a small sign added to a letter to alter pronunciation or to distinguish between similar words. A diacritical mark can appear above or below a letter or in some other position. Its main usage is to change the phonetic value of the letter to which it is added. The diacritic mistake is not often made in English while it is more likely to happen in multi-lingual collections (especially in European language). In Figure 47, it is clear that the name in the first row is same as the name in the second row. The computer programs can identify this problem if they have appropriate conversion tables. According to the Levenshtein distance, the edit distance between these strings is two. If the string contains three or more diacritic characters, Mat will not be able to detect this problem.



Original Text	Source Link
Rasovic, A.	http://hdl.handle.net/10092/392
Rašović, A.	http://hdl.handle.net/10092/551

Figure 47 : A screenshot of a potential duplicate list



Rašović, A.
Rabczuk T.
Rabczuk, T.
Raffensperger, J.F.
Ramos, G.
Rance, B.D.
Rangiheuea, T.
Rasovic, A.

Figure 48 : A screenshot of an author list – two variations of "Rasovic"

5.3 Sample Collection Analysis

The creation of digital resources has been increasing rapidly in recent years. Metadata elements should be used correctly in order to maintain the quality of repositories. This section describes a study of how Data Providers use the simple Dublin Core Metadata to index their collections. We analyse six institutional repositories that have exposed the metadata through the OAI-PMH. "The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is the protocol that facilitates digital library interoperability and cross-domain resource discovery" (Lagoze et al., 2002). There are two types of participants in the OAI-PMH framework: Data Providers and Service providers. Data Providers expose the metadata which can be harvested by the Service Providers. In this study, institutional repositories are the Data Providers and Greenstone is the Service Provider.

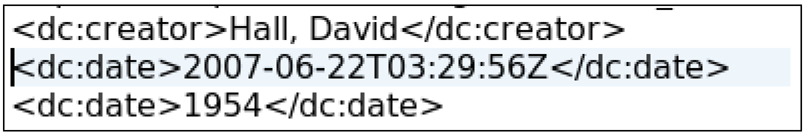
OAI defines a mechanism for Data Providers to expose the their metadata records and mandates the Data Providers to map their metadata records to the unqualified Dublin Core (i.e. simple Dublin Core). Although Dublin Core Metadata Set defines 15 elements, the majority of Data Providers only use small part of the elements. Ward (2002) found that the general population of OAI-compliant collections had only eight DC elements defined. In this study, we would like to examine the degree to which each repository uses the unqualified Dublin Core Metadata Set.

We used Greenstone as the Service Provider to harvest metadata records between 21/05/2008 and 27/05/2008 from six Data Providers and the total 15,661 metadata records were harvested. Table 6 shows the characteristics of those six collections (the URLs are shown in Appendix C). Each Data Provider uses at least nine distinct metadata elements to describe their records.

Table 7 shows summary statistics for 14 Dublin Core elements. The elements are ordered by their total occurrences (frequency) in the metadata. The total number of metadata records harvested from six institutional repositions is 15,661. The average number of metadata records per repositories is 2610. The top five elements used in this study are: Subject (15.9 percent), Date (13.2 percent), Format (12.7 percent), Creator (11.7 percent), and Identifier (10.8 percent). There are four elements used less than one percent in metadata records. Surprisingly, no repositories use the DC coverage element in any metadata records. All repositories include the "title" element in their collections. There is an average of 17 DC elements used per record. It suggests that several elements tend to occur more frequently.

The statistics shown in Table 7 is not consistent with Ward's findings. Ward (2002) reported that the most commonly used element in the 910,919 metadata records was "Creator" (21.5 percent) followed by "Identifier" (17.2 percent), "Title" (11.4 percent), "Date" (11.1 percent) and "Type" (10.7 percent). Ward also found that Data Providers used an average of eight elements per record. The difference can be due to greater diligence by metadata contributors or some other factors such as large data set and types of repositories (Jordan, 2006).

The unqualified Dublin Core is a simple yet effective element set for describing resources. The advantage of it is easy to be harvested by Service Providers via the OAI-PMH. However, the disadvantage is that it is extreme simple, so mapping from other richer metadata records can lead to loss of information. For example, the qualified Dublin Core defines several types of Date such as date of creation and date of issue. When the qualified Dublin Core is mapped to the unqualified Dublin Core, it will cause ambiguity between these two dates. Because unqualified Dublin Core cannot distinguish between date of creation and date of issue, they will both be represented as "dc.date" in unqualified Dublin Core. Figure 49 shows an example of the ambiguity in unqualified Dublin Core.



```
<dc:creator>Hall, David</dc:creator>
<dc:date>2007-06-22T03:29:56Z</dc:date>
<dc:date>1954</dc:date>
```

Figure 49 : A screenshot of an OAI record

	Collection 1	**Collection 2**	**Collection 3**	**Collection 4**	**Collection 5**	**Collection 6**
Total number of records	5964	6015	1964	667	499	276
Size of repository	large	large	medium	small	small	Small
Type of resources described	Pre-prints, papers, reports, data sets …	pre-prints, papers, thesis, presentations post-print …	Papers Thesis	Book chapters, theses, discussion papers, Journal	Journal, Conference papers Working papers	Journal, Conference papers Working papers
Number of DC element used	12	13	12	13	9	9

Table 6 : Statistical characteristic of six repositories

Summary of Metadata Records					
DC Element	Total times of element used	Each element as a % of the Total times of Element used (261,261)	Average times used per record (15,385)	Number of records containing element	% of total records containing element
Subject	41,455	15.9	2.7	12,357	80
Date	34,132	13.2	2.2	15,351	99
Format	33,227	12.7	2.2	10,744	70
Creator	30,528	11.7	2.0	12,545	81
Identifier	28,327	10.8	1.8	15,368	99
Type	15,474	5.9	1.0	13,766	89
Title	15,399	5.9	1.0	15,385	100
Language	14,814	5.7	1.0	12,345	80
Relation	12,805	4.9	0.8	9,115	59
Publisher	11,288	4.3	0.7	11,237	73
Contributor	9,554	3.6	0.6	8,959	58
Rights	8,641	3.3	0.6	7,648	49
Source	5,617	2.1	0.4	5,554	36
Coverage	0	0	0	0	0
Total	311,027	100	17	---	---

Table 7 : Summary of Metadata Records

Analysis of the data

Mapping of qualified Dublin Core to simple Dublin Core produces several problems of ambiguity. For our analysis, we will examine the values of the format, identifier, type, and date.

Format

The Dublin Core Metadata Initiative (DCMI) defines format as the file format, physical medium, or dimensions of the resource (Dublin Core Metadata Initiative, 2008). There are four collections using the format element more than once and two common types of variations can be derived from them.

Firstly, metadata records use format to describe their digital formats such as text/xml, text/html, or pdf.

Secondly, it is used to describe the size of files. In one of the sample collection, there are 7400 unique values for format element. But over 95 percent of them are filesize and the rest are their digital formats.

Identifier

The DCMI defines identifier as an unambiguous reference to the resource within a given context such as ISBN, URL, and ISSN (Dublin Core Metadata Initiative, 2008). The identifier should be a unique ID for the resource and used once. However, small numbers of records in the sample collection do not use the identifier elements.

The most common type of identifier is URL which links to the destination of the documents in the local repositories. Small numbers of identifiers contain the locally-derived internal numbers rather than the URL. In the analysis of our sample records, no collection maintained a consistent one-to-one mapping between the identifier element and the resources. As shown in Table 7, every collection assigns more than one Identifier element to metadata records. In our analysis, most collections assign one internal identifier and one URL to metadata records.

Type

The DCMI defines type as the nature, genre, purpose, and function of the resource (Dublin Core Metadata Initiative, 2008). It is important not to confuse "type" with "subject" or "format". In our analysis, three collections use the type element to describe both type and format. For example, some repository managers use this element to describe the digital format of the resources. They also use different words to describe the same object such as book section, book chapter, and book.

Date

The DCMI defines data as a point or period of time associated with an event in the lifecycle of the resource and suggests using the WC3DFT profile to describe the date element (Dublin Core Metadata Initiative, 2008). The WC3DFT profile defines six levels of granularity in the date and time.

- Year: YYYY (e.g. 1997)
- Year and month: YYYY-MM (e.g. 1997-07)
- Complete date: YYYY-MM-DD (e.g. 1997-07-16)
- Complete date plus hours and minutes:

  YYYY-MM-DDThh:mmTZD (e.g. 1997-07-16T19:20+01:00)
- Complete date plus hours, minutes and seconds:

  YYYY-MM-DDThh:mm:ssTZD (e.g. 1997-07-16T19:20:30+01:00)
- Complete date plus hours, minutes, seconds and a decimal fraction of a second

  YYYY-MM-DDThh:mm:ss.sTZD

  (e.g. 1997-07-16T19:20:30.45+01:00)

As mentioned earlier, the simple Dublin Core metadata elements are flexible. In our analysis, many metadata records in the sample collections use the date element more than twice. As shown in Figure 49, that metadata record contains two date elements. That resource was created in 1954 and imported into the repository in 2007.

The majority of records use the complete date plus hours and minutes date format. This timestamp is typically assigned by computer software when the resource is submitted or imported to a collection.

Element	**Collection 1**	**Collection 2**	**Collection 3**	**Collection 4**	**Collection 5**	**Collection 6**
Subject	2.7	4.7	1.6	2.1	2.8	5.0
Date	1.0	3.2	3.0	1.0	3.0	3.9
Format	1.0	4.5	2.5	1.2	3.4	1.0
Creator	3.0	2.6	1.1	1.6	1.7	2.8
Identifier	2.0	1.9	1.1	2.0	2.4	2.3
Type	1.1	1.0	1.0	2.0	2.2	1.0
Title	1.0	1.0	1.0	1.0	1.0	1.0
Language	1.0	1.4	1.4	0.0	1.0	0.0
Relation	1.4	1.9	1.0	1.6	1.3	1.0
Publisher	1.0	1.0	1.0	1.0	1.0	1.0
Contributor	1.0	1.1	1.5	1.0	1.7	0.0
Rights	1.0	1.3	1.1	1.0	1.0	0.0
Source	1.0	0.0	1.0	0.0	0.0	0.0
Coverage	0.0	0.0	0.0	0.0	0.0	0.0

Table 8 : Average usage of Dublin Core metadata elements in each repository

5.4 User Feedback

We used an online survey and interviews with repository managers to gather feedback (see appendix A). Most feedback received was from repository managers, though some were still planning or developing their repositories. Generally, the remote surveys have been partially successfully in eliciting feedback for improvement, and face-to-face traditional usability think-aloud has been more useful (Nichols and Paynter et al., 2008). Most feedback was generally positive and has been arranged into logical groups in this section.

5.4.1 Statistical Report

Participant A is a repository manager and has tried to use other metadata quality tools to increase the visibility of the records in his repository. Participant A thinks the Mat visualisation tool is useful when finding the incomplete records.

> *I see more ways to view lists of each specific metadata element than last time I used MAT - really great! Potential duplicate list is fantastic, especially for subject and author.*

He thinks it would be useful to run regular reports of the New Zealand repositories for ease of access.

Participant B is a repository manager and looking for a tool to increase to the metadata quality for his library. He mentioned that his native repository software cannot isolate problem records like Mat. Mat allows him to see which Dublin Core fields are incomplete, as well as highlighting problems such as white space and punctuations at the beginning of fields.

> *A list of incomplete records that are hyperlinked is fantastic because I do not have to go to the repository to search for them. It makes improving the quality of our metadata easier.*

Participant B thinks Mat will be more useful if it has the ability to apply to other metadata schemas such as MARCXML.

Participant C is a repository manager and is now in charging of setting up a new institutional repository. His first repository contains some bad data and caused lots of trouble for him. He is looking for a useful tool that could evaluate metadata quality.

> *I can see where that would be really useful, because you crosswalk everything into DC, and it would be great to see where that crosswalking is failing*

Participant C suggests that the metadata validation could be added into the next version. For example, the tool can determine whether the appropriate data have been entered for a particular record.

5.4.2 Visualisation Tool

Participant D is a repository manager and has just discovered Mat recently. His repository is about research and innovation in social services.

> *I found and used this tool last week and found it very useful for exploring our own repository. I used it to export a list of our metadata for analysis & re-use in the customised search interface.*

Participant D hopes to see this tool in the next version of Greenstone.

Participant E describes Mat as incredibly useful and very exciting. Participant E thinks Mat would be most useful in the early stages of repository development.

> *I especially like the graphical representation because the completeness of metadata is a mark of quality. It is useful to be able to examine what has being harvested by OAI harvesters.*

Participant E also thinks the system will be more useful to work with other metadata schemes.

5.4.3 Entire System

Participant F is a repository manager and used Mat to generate a list of URLs to check metadata records.

> *I see you have discovered our problems; the author hofig contained an encoded character and totally misread the index position. I have changed it so it can be searched and appears in the correct position in the index.*

Participant G is an institutional repository manager. She was introduced to Mat in January 2008 and found it was very useful. Since Mat has been upgrade to second version, Participant G was re-interviewed in July 2008. She also noticed and corrected a number of missing abstracts and incorrect copyright fields that she would not otherwise have noticed.

> *I noticed the interface had been update to allow me to directly click through to a record, and that that was extremely useful to me.*

Her institution is running a performance and development program with rewards that requires employees to have measurable objectives, she and her colleagues are planning on using Mat to determine the percentage completeness of metadata sets in both the image and research repositories as performance goals.

5.5 Summary

This chapter has discussed six common types of mistakes/errors in metadata element values. Some errors are very hard to detect without using software applications. For example, there is very little difference between “Department” and “Depatment”. It is very easy to miss this kind of errors for repository managers who are responsible for maintaining large collections. Repository managers can use this tool to analyse their collections for finding errors in metadata element values.

In section 5.3, the usage of the Dublin Core Metadata Element Set by different institutions was described. Most repositories used at least nine Dublin Core metadata elements to describe their resources. Due to the flexibility of simple Dublin Core, no collection maintained a one-to-one mapping between records and resources. In section 5.4, we presented qualitative feedback in order to examine the usability of Mat. Most feedback was positive and encouraging.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

As the amount of information increases in digital libraries, it becomes very difficult for repository managers to maintain the quality of their collection. In this research we have studied the problems of metadata quality and constructed a practical tool to aid metadata quality assessment.

The main contributions of this research are:

- Building a metadata analysis tool that is integrated with Greenstone.
- Creating the first public web-based quality analysis tool for digital collections.
- Exploring the feasibility of web-based visualisation for quality analysis.
- Generating qualitative feedback to clarify requirements for metadata tools.
- Identifying types of errors in current institutional repositories.

We constructed the tool by extending the existing Greenstone system. However, during the development of Mat, we discovered several bugs in the Greenstone code. Previously, nobody had tried to build hundreds of collections with Greenstone and these problems could be only discovered if there are more than 50 or 60 collections on a Greenstone server. Once, this problem was fixed and the stability of Mat increased significantly. This example illustrates the dependence of Mat on the underlying of Greenstone system.

We have shown that our analysis tool, despite its exploratory nature, is practical and useful for managers of digital collections.

6.2 Future Work

As mentioned in chapter 5, repository managers have suggested many improvements to Mat. Four areas of future work are listed below:

6.2.1 Mat Improvements

Integration with Greenstone3

This tool was original designed to be a part of the Greenstone Digital System but it is not in the current distribution. Once Mat has been integrated with Greenstone3 then users can setup their own local analysis tool. It should also be easier for users to customise the appearance of their own version of Mat.

Express profiles – Rules for Each Metadata Element

One repository manager suggested that a metadata quality tool should be able to determine whether the appropriate data have been entered for a particular record. For example, a date element should not accept an Email address as its metadata value. To address this problem, we plan to add content rules (e.g. regular expression) to validate metadata records.

Using Different Shading Techniques to Improve Visualisation

Currently, we use a blue rectangle to indicate the presence of an element and this technique cannot show the number of occurrences of that element. In the future, we are going to improve the visualisation tool by applying different colours or shades to represent the occurrences of metadata elements.

Improving the Edit Distance

The technique we used to catch metadata errors is the Levenshtein distance and it is useful for detecting typographical errors. Although it is a powerful algorithm, some errors are beyond its scope. It could not catch abbreviation problems. For example, "gym" is the abbreviation of "gymnasium" and the edit distance for these two words is six. Because the default threshold is two, these two words are considered different. This type of error is commonly seen in authors' names. To increase the accuracy of the potential duplicate list, we plan to modify this algorithm to detect these types of errors.

6.2.2 Greenstone3 improvements

Metadata Scheme

As mentioned in chapter 5.4, many repository managers suggested that Mat will be more useful to work with other metadata schemes. However, Greenstone must be updated in order to achieve this goal. Greenstone is responsible for harvesting metadata records and building collections. However, the current version of Greenstone does not work well with other schemes such as METS (Metadata Encoding and Transmission Standard, 2008) and MODS (Metadata Object Description Schema: MODS, 2008).

Incremental harvesting

Incremental harvesting means a harvester only needs to retrieve metadata records from the last harvest date for a repository. This technique could be used reduce the amount of harvesting workload. If Greenstone is upgraded to support incremental harvest then Mat will be significantly faster.

References

Ahlberg, C., Shneiderman, B. 1994. Visual Information Seeking: Tight coupling of dynamic query filters with starfield displays. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Celebrating Interdependence*, 313-317.

Bainbridge, D., Don, K. J., Buchanan, G. R., Witten, I. H., Jones, S. R., Jones, M. and Barr, M. I. 2004. Dynamic digital library construction and configuration. *Proceedings of the Eighth European Conference on Research and Advanced Technology for Digital Libraries (ECDL'04),* LNCS 3232, 1-13. Springer-Verlag, Berlin.

Barton, J., Currier, S., and Hey, J. 2003. Building Quality Assurance into Metadata Creation: an Analysis based on the Learning Objects and e-Prints Communities of Practice. *Proceedings of 2003 Dublin Core Conference: Supporting Communities of Discourse and Practice - Metadata Research and Applications*, 39-48.

Beall, J. 2005. Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information*, 6(3). http://jodi.tamu.edu/Articles/v06/i03/Beall/

Beall, J. and Kafadar, K. 2007. Measuring Typographical Error's Impact on Retrieval in Bibliographic Databases. *Cataloging and Classification Quarterly* 44(3/4), 197-211.

Bruce, T. R. and Hillmann, D. I. 2004. The continuum of metadata quality: defining, expressing, exploiting. In *Metadata in Practice*, American Library Association, Chicago, IL. 238-256.

Buchanan, G., Bainbridge, D., Don, K., and Witten, I. H. 2005. A New Framework for Building Digital Library Collections. *Proceedings of the Fifth ACM/IEEE-CS Joint Conference on Digital Libraries*, 23-31. (JCDL, 2005)

Bui, Y. and Park, J-R, 2006. An assessment of metadata quality: A case study of the national science digital library metadata repository. http://idea.library.drexel.edu/bitstream/1860/1600/1/2007021006.pdf

Davis, J. R. and Lagoze, C. 2000. NCSTRL: design and deployment of a globally distributed digital library. *Journal of the American Society for Information Science*, 51(3), 273-280.

Dublin Core Collection Description Working Group. 2004. Dublin Core Collection Description Proposed Term: Provenance. Retrieved 21 July, 2008 from http://www.ukoln.ac.uk/metadata/dcmi/collection-provenance/

Dublin Core Metadata Initiative. 2008. Dublin Core Metadata Element Set, Version 1.1. Retrieved 25 June, 2008 from http://dublincore.org/documents/dces/

Duval, E. and Ochoa, X. 2006. Towards automatic evaluation of metadata quality in digital repositories, In *Advances in Conceptual Modeling-Theory and Practice, ER 2006 Workshops BP-UML*, Springer, 372-381.

Efron, M. 2007. Metadata use in OAI-Compliant Institutional Repositories.
Journal of Digital Information, 8(2).
http://journals.tdl.org/jodi/article/view/196/169

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. 2007. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1). http://www.cs.purdue.edu/homes/ake/pub/survey2.pdf

Fox, E. A. 1999. The Digital Libraries Initiative - Update and Discussion. *Bulletin of the America Society of Information Science*, 26(1), 7-27.

Gasser, L. and Stvilia, B. 2001. A new framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), 1720-1723.

Greenberg, J., Pattuelli, M. C., Parsia, B., and Robertson, W. D. 2001.
Author-generated Dublin Core metadata for web resources: A baseline study in an organization. *Proceedings of the International Conference on Dublin Core and Metadata Applications 2001*. http://journals.tdl.org/jodi/article/view/jodi-39/45

Guy, M., Powell, A., and Day, M. 2004. Improving the Quality of Metadata in Eprint Archives. *Ariadne* Issue 38. http://www.ariadne.ac.uk/issue38/guy/

Hillmann, D. I. 2005. *Using Dublin Core*. Retrieved 25 June, 2008 from http://dublincore.org/documents/usageguide/

Hillmann, D. I. and Dushay, N. 2003. Analyzing metadata for effective use and re-use. *Proceedings of the International Conference on Dublin Core and Metadata Applications*. http://www.siderean.com/dc2003/501_Paper24.pdf

Hughes, B. 2004. Metadata quality evaluation: experience from the open language archives community. *Proceedings of the Seventh International Conference on Asian Digital Libraries ICADL2004)*, 320-329.

International Development Research Centre: International Development Research Centre. n.d. International Development Research Centre: International Development Research Centre. Retrieved 25 July, 2008 from http://www.idrc.ca/en/ev-1-201-1-DO_TOPIC.html

Jordan, M. 2006. The CARL metadata harvester and search service. *Library Hi Tech*, 24(2), 197-210.

Kahn, R. and Wilensky, R. 1995. A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2), 115-123.

Lagoze, C., Van de Somple, H., Nelson, M., and Warner, S. 2002. The Open Archives Initiative Protocol for Metadata Harvesting. Retrieved 23 June, 2008 from http://www.openarchives.org/OAI/openarchivesprotocol.html

Lesk, M. 1997. *Practical Digital Libraries*. Morgan Kaufmann, San Francisco, CA.

Marks, L., Hussell, J. A. T., McMahon, T. M., and Luce, R. E. 2005. ActiveGraph: A Digital Library Visualisation Tool. *International Journal on Digital Libraries*, 5(1), 57-69.

Marzal, A., and Vidal, E. 1993. Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9), 926-932.

Moen, W. E., Stewart, E. L., and McClure, C. R. 1997. The Role of Content Analysis in Evaluating Metadata .for the U.S Government Information Locator Service (GILS): Result from an Exploratory. http://www.unt.edu/wmoen/publications/GILSMDContentAnalysis.htm

Moen, W. E., Stewart, E. L., and McClure, C. R. 1998. Assessing metadata quality: findings and methodological considerations from an evaluation of the US Government Information Locator Service (GILS). *Proceedings of the Research and Technology Advances in Digital Libraries Conferences*, 246-255.

Moen, C and McCluren, C. 1997. An Evaluation of the Federal Government's Implementation of the Govermenet Information Locator Service (GILS). http://www.gpoaccess.gov/gils/gils-eval.pdf

Metadata Object Description Schema: MODS (Library of Congress). 2008. Metadata Object Description Schema: MODS (Library of Congress). Retrieved 25 August, 2008 from http://www.loc.gov/standards/mods/

Metadata Encoding and Transmission Standard (METS). 2008. Metadata Encoding and Transmission Standard (METS). Retrieved 25 August, 2008 from http://www.loc.gov/standards/mods/

Najjar, J., Ternier, S., and Duval, E. 2003. The actual use of metadata in Ariadne: An empirical analysis. *Proceedings of the Third Annual ARIADNE Conference*, 1-6.

Najjar, J., Ternier, S., and Duval, E. 2004. User Behavior in Learning Object Repositories: An Empirical Analysis. *Proceedings of the ED-MEDIA 2004 World Conference on Educational Multimedia, Hypermedia and Telecommunications,* 4373-4379.

Nichols, D. M., Paynter, G. W., Chan C-H., Bainbridge, D., McKay, D., Twidale, M.B. and Blandford, A. 2008. *Metadata Tools for Institutional Repositories* (Working paper 2008/10). Hamilton, New Zealand: University of Waiakto, Department of Computer Science.

Nichols, D. M., Chan, C-H., Bainbridge, D., McKay, D. and Twidale, M. B. 2008. Metadata Tools for Institutional Repositories. *Proceedings of the Eighth ACM/IEEE-CS Joint Conference on Digital Libraries*, 201-210. (JCDL 2008)

Navarro, G. 2001. A guided tour to approximate string matching. *ACM Computing Surveys,* 33(1), 31-88.

OAI：Open Archives Initiative. 2008. OAI：Open Archives Initiative. Retrieved 27 June, 2008 from http://www.lib.tku.edu.tw/esource/scholar/project/OAI.htm

Sánchez, J. A., Quintana, M. G., and Razo, A. 2007. A Star-fish: starfields+fisheye visualisation and its application to federated digital libraries. *Proceedings of the Third Latin American Conference on Human-Computer Interaction*. http://clihc.org/2007/papers/StarFish_ID38_longpaper.pdf

Sánchez, J. A., Twidale, M. B., Nichols, D. M., and Silva, N. 2005. Experiences with starfield visualizations for analysis of library collections. *Proceedings of the Visualisation and Data Analysis Conference* (VDA 2005), 215-225.

Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M. B. and Cole, T. W. 2005. Is quality metadata 'shareable' metadata? The implications of local metadata practices for federated collections. *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries,* 223-237.

Stvilia, B., Gasser, L., Twidale, M. B., Shreeves. S. L., and Cole, T. W. 2004. Metadata quality for federated collections. *Proceedings of the Ninth International Conference on Information Quality*.
http:://www.isrl.uiuc.edu/~stvilia/papers/iciq_144_final_v1.pdf

Ward, J. 2002. A quantitative analysis of unqualified Dublin Core metadata element set usage within data providers registered with the open archives initiative.
Proceedings of the Third ACM/IEEE-CS Joint Conference on Digital Libraries, 315-317.

White, T. 2004. Can't beat Jazzy – Introducing the Java platform's Jazzy new spell checker API. Retrieved 29 June, 2008 from
http://www.ibm.com/developerworks/java/library/j-jazzy/

Wilson, A. J. 2007. Toward releasing the metadata bottleneck - a baseline evaluation of contributor-supplied metadata. *Library Resources and Technical Services*, 51(1), 16–28.

Witten, I. H. and Bainbridge, D. 2003. *How to Build a Digital Library*. Morgan Kaufmann, San Francisco, CA.

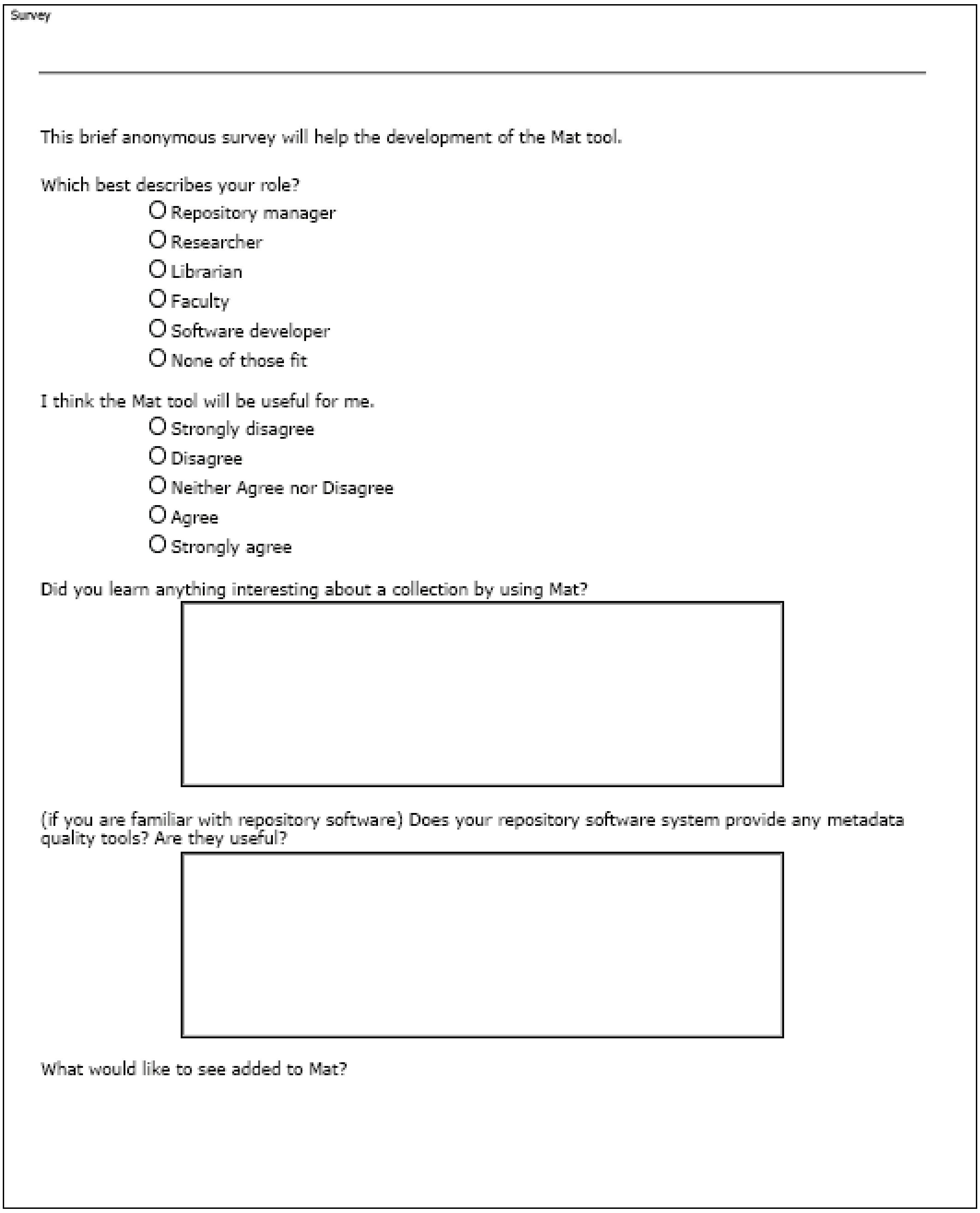
Zeng, M. L., Subrahmanyam, B., and Shreve, G. M. 2004. Metadata quality study for the National Science Digital Library (NSDL) Metadata Repository. *Proceedings of the Seventh International Conference on Asian Digital Libraries, ICADL 2004*, 339-340.

Appendices

Appendix A: User Study Material

This Appendix contains a questionnaire, a research consent form, a description of the experiment, and confirmation of ethical approval.

Questionnaire



Survey

This brief anonymous survey will help the development of the Mat tool.

Which best describes your role?

○ Repository manager
○ Researcher
○ Librarian
○ Faculty
○ Software developer
○ None of those fit

I think the Mat tool will be useful for me.

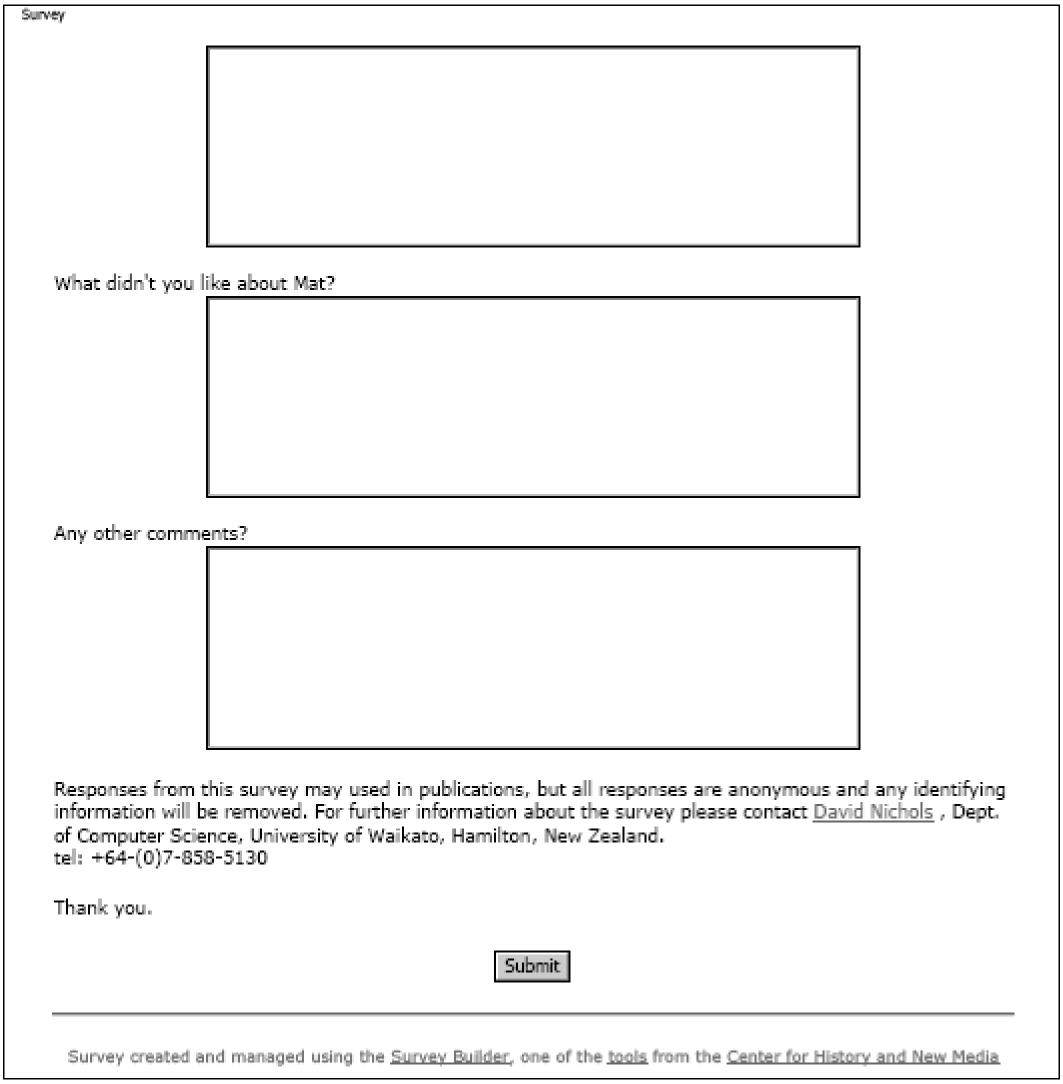
○ Strongly disagree
○ Disagree
○ Neither Agree nor Disagree
○ Agree
○ Strongly agree

Did you learn anything interesting about a collection by using Mat?

(if you are familiar with repository software) Does your repository software system provide any metadata quality tools? Are they useful?

What would like to see added to Mat?

Questionnaire Page 1



What didn't you like about Mat?

Any other comments?

Responses from this survey may used in publications, but all responses are anonymous and any identifying information will be removed. For further information about the survey please contact David Nichols , Dept. of Computer Science, University of Waikato, Hamilton, New Zealand.
tel: +64-(0)7-858-5130

Thank you.

Survey created and managed using the Survey Builder, one of the tools from the Center for History and New Media

Questionnaire Page2

The University of Waikato · School of Computing and Mathematical Sciences

Research Consent Form

> *This consent form, a copy of which has been given to you, is only part of the process of informed consent. It should give you the basic idea of what the research is about and what your participation will involve. If you would like more detail about something mentioned here, or information not included here, please ask. Please take the time to read this form carefully and to understand any accompanying information.*

Research Project Title

Evaluation of a metadata analysis tool

Researchers

Dr David Nichols, Eric Chan, Dana McKay

Experiment Purpose

The purpose of this experiment is to evaluate a prototype metadata analysis tool and to generate ideas for future improvements of the tool.

Participant Recruitment and Selection

Repository managers, librarians and other managers of digital collections are recruited via online publicity (blogs, email, mailing lists etc.) and personal contacts.

Procedure

This session should not require more than about an hour of your time. You will be asked to use the online tool to evaluate online digital collections and we may ask questions as you proceed. You will also be invited to fill in an anonymous web questionnaire about your experience with the tool. None of the tasks are a test – the objective is to find out how to improve the tool and make it more useful for repository managers.

Data Collection

We may make notes as you use the tool. In the case of a telephone interview we may record the telephone conversation as a digital audio file. You will also be invited to fill in an anonymous web questionnaire about your experience with the tool.

Data Archiving/Destruction

The audio recordings will be transcribed and the files deleted. Transcripts, interview notes and web survey responses will be stored in the SCMS Data Archive under the control of the School Ethics Committee and the School Manager (Dean's Office). They will be destroyed on 31/1/2013.

Confidentiality

Confidentiality and participant anonymity will be strictly maintained. No names or other identifying characteristics will be stated in the final or any other reports. References to identifying information such as specific collections will be removed if quotations are used in publications.

Likelihood of Discomfort

There is no likelihood of discomfort or risk associated with participation.

Researchers

David Nichols is a senior lecturer in the Computer Science Department at the University of Waikato. This study will contribute to his research on digital libraries and metadata quality.

David can be contacted in room G.2.08 of the School of Computer and Mathematical Sciences building at the University of Waikato. His phone number is +64(7)8585130and his email address is dmn@cs.waikato.ac.nz.

Eric Chan is a Masters student working on metadata quality in digital libraries. His supervisor is Dr David Nichols. Email: cc108@waikato.ac.nz

Dana McKay is a PhD student working on information retrieval from digital collections. Email: dana@cs.waikato.ac.nz

Finding out about Results

The Participants can find out the results of the study by contacting the researcher after Jun 1, 2008.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in the research project and agree to participate as a participant. In no way does this waive you legal rights nor release the investigators, sponsors, or involved institutions from their legal and professional responsibilities. You are free to not answer specific items or questions in interviews or on questionnaires. You are free to withdraw from the study at any time without penalty. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please contact the researcher.

__________________________ __________________

Participant Date

__________________________ __________________

Investigator/Witness Date

A copy of this consent form has been given to you to keep for your records and reference.

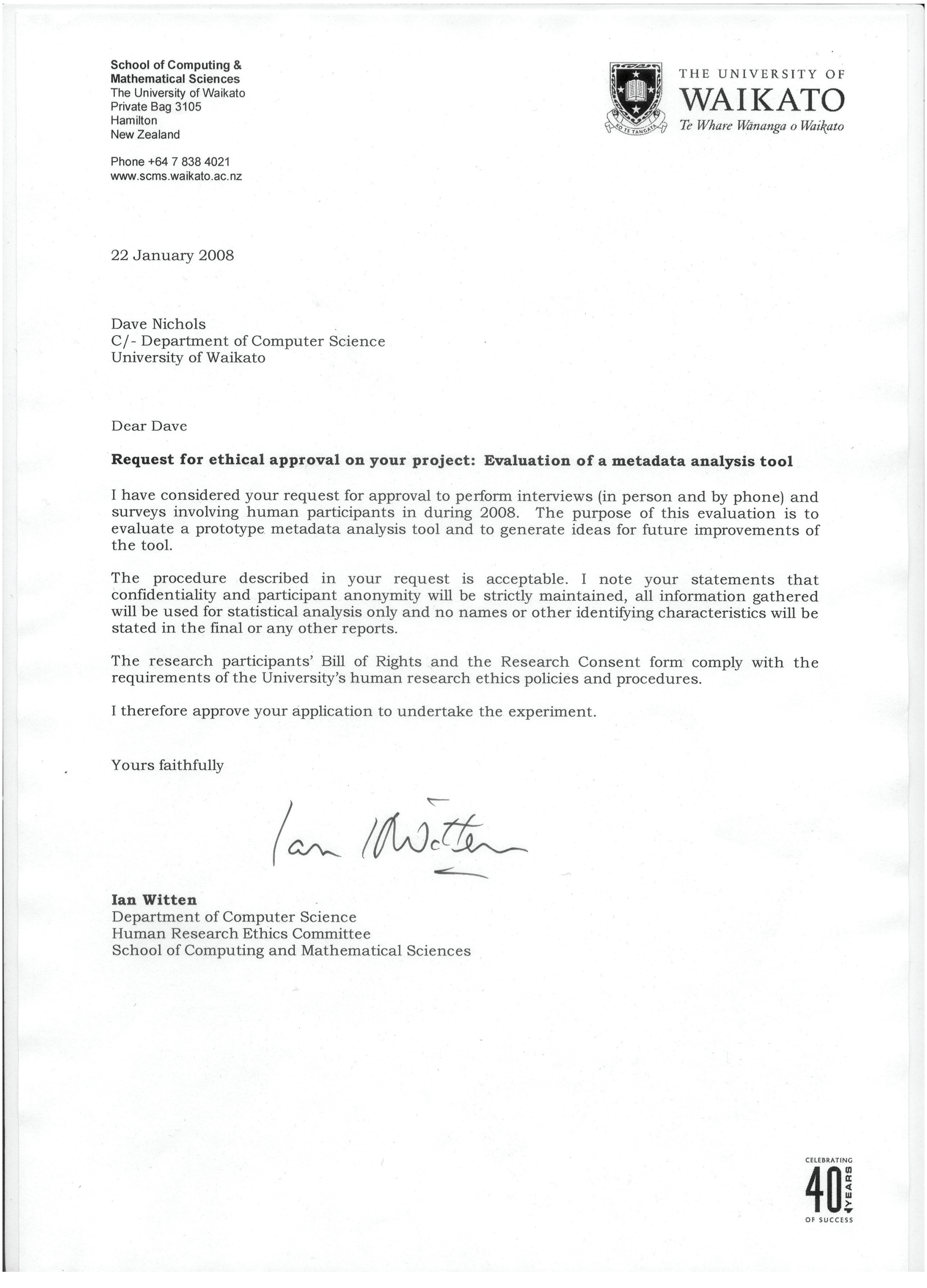
Study Description

The in-person task description is on the following page.

Prior to a phone-based study the study and task descriptions, consent form and bill of rights will have been emailed to the participant and email consent obtained. At the start of the phone study verbal consent will be asked for again to confirm the participant's involvement.

For a phone-based study the participant will be asked to follow the in-person task description. Either during, or following the study, an experimenter will ask the participant brief verbal questions about their experiences using the tool.

The anonymous web survey and screenshots of the tool are on the pages following the in-person task description.



School of Computing &
Mathematical Sciences
The University of Waikato
Private Bag 3105
Hamilton
New Zealand

Phone +64 7 838 4021
www.scms.waikato.ac.nz

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

22 January 2008

Dave Nichols
C/- Department of Computer Science
University of Waikato

Dear Dave

Request for ethical approval on your project: Evaluation of a metadata analysis tool

I have considered your request for approval to perform interviews (in person and by phone) and surveys involving human participants in during 2008. The purpose of this evaluation is to evaluate a prototype metadata analysis tool and to generate ideas for future improvements of the tool.

The procedure described in your request is acceptable. I note your statements that confidentiality and participant anonymity will be strictly maintained, all information gathered will be used for statistical analysis only and no names or other identifying characteristics will be stated in the final or any other reports.

The research participants' Bill of Rights and the Research Consent form comply with the requirements of the University's human research ethics policies and procedures.

I therefore approve your application to undertake the experiment.

Yours faithfully

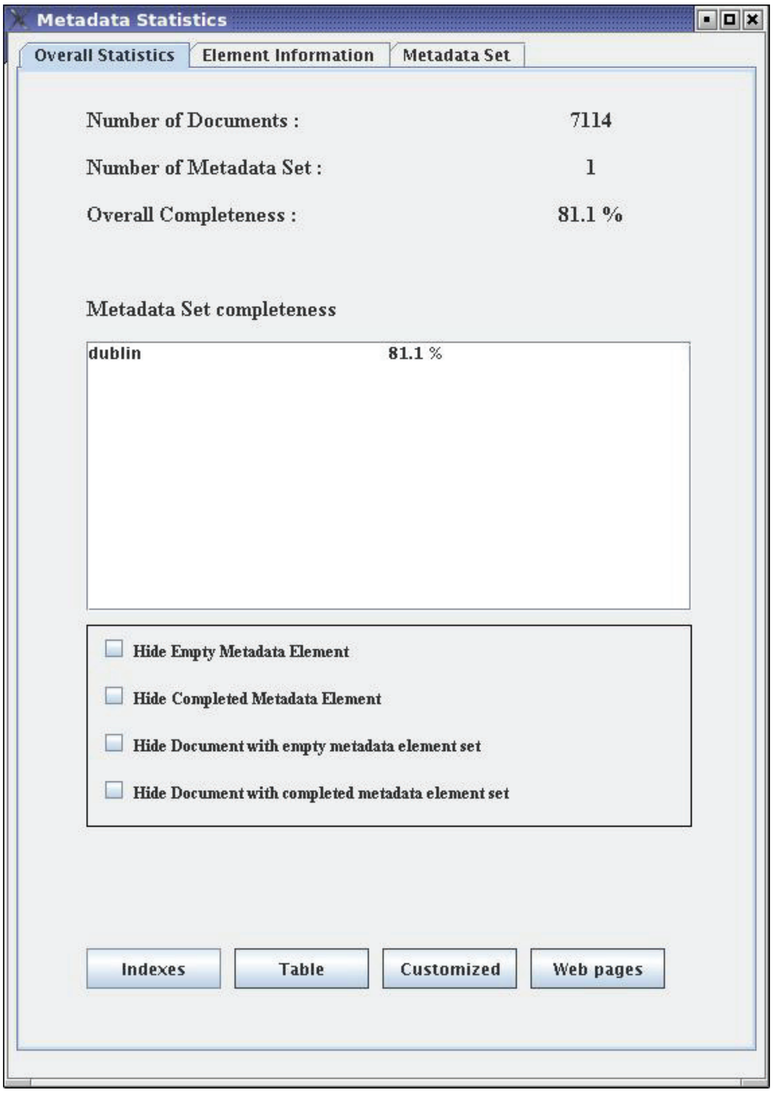
Ian Witten

Ian Witten
Department of Computer Science
Human Research Ethics Committee
School of Computing and Mathematical Sciences

CELEBRATING 40 YEARS OF SUCCESS

Appendix B: Additional Screenshots

This appendix contains a serious of screenshot of Mat (both Java and web version).



Number of Documents : 7114

Number of Metadata Set : 1

Overall Completeness : 81.1 %

Metadata Set completeness

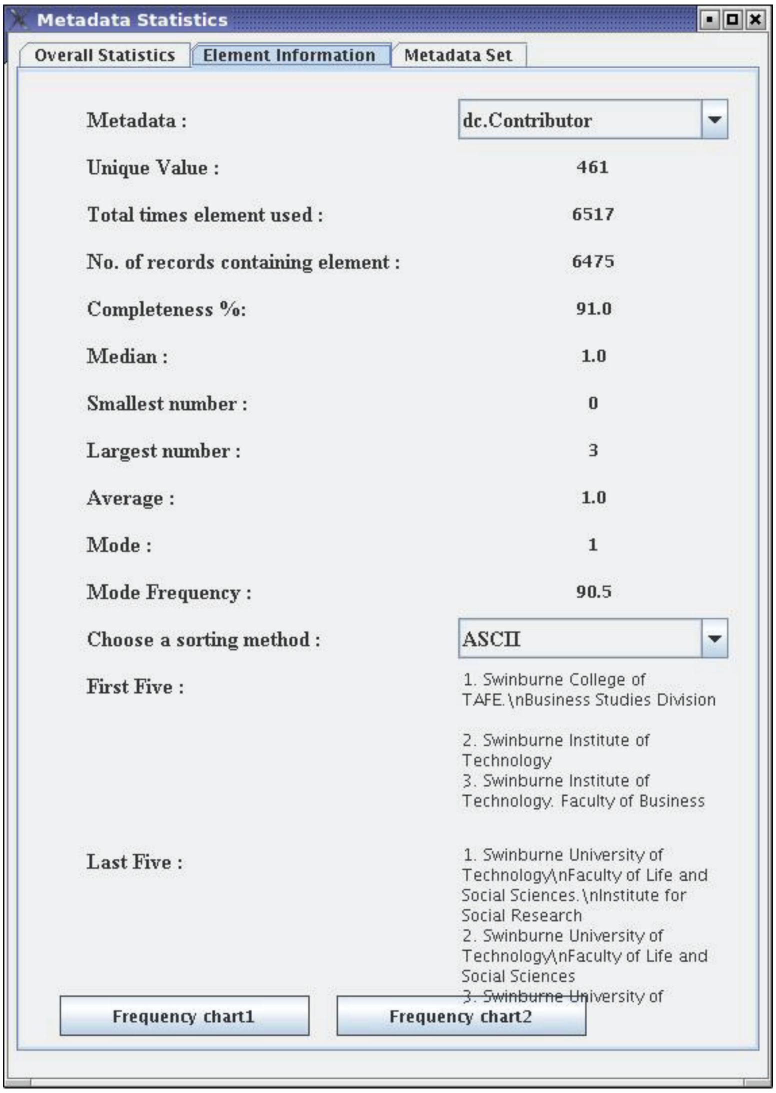
dublin	81.1 %

☐ Hide Empty Metadata Element

☐ Hide Completed Metadata Element

☐ Hide Document with empty metadata element set

☐ Hide Document with completed metadata element set



Metadata : dc.Contributor

Unique Value : 461

Total times element used : 6517

No. of records containing element : 6475

Completeness %: 91.0

Median : 1.0

Smallest number : 0

Largest number : 3

Average : 1.0

Mode : 1

Mode Frequency : 90.5

Choose a sorting method : ASCII

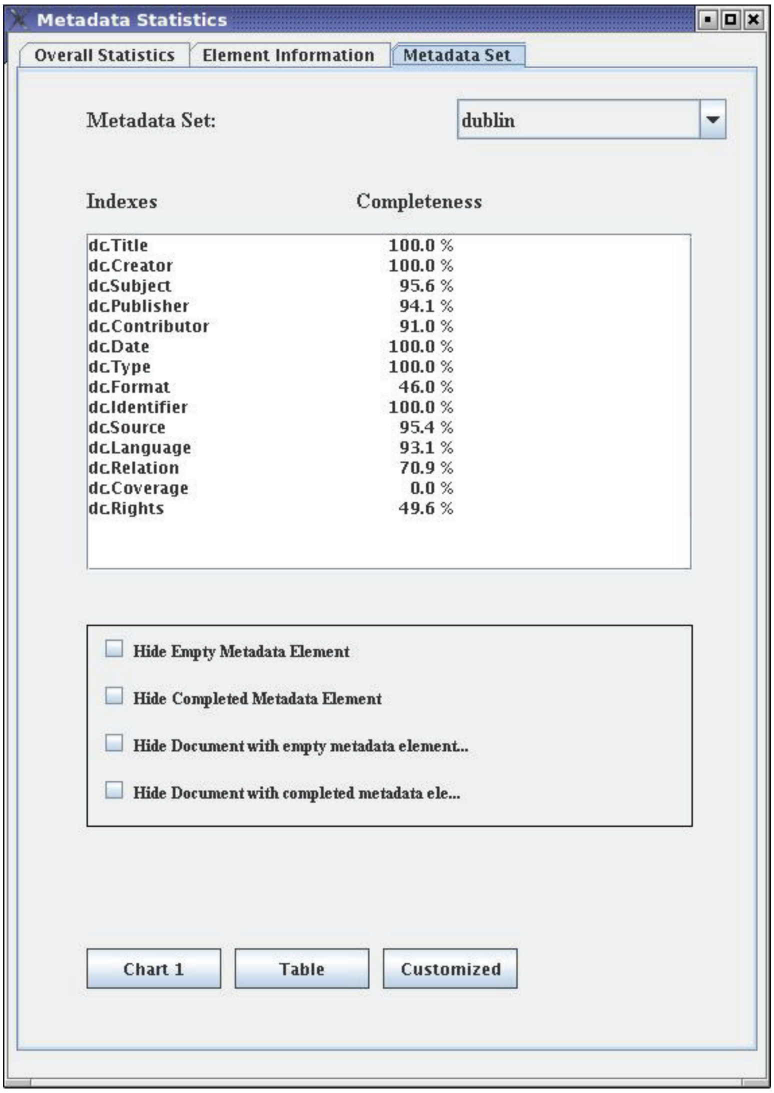
First Five :

1. Swinburne College of TAFE.\nBusiness Studies Division

2. Swinburne Institute of Technology
3. Swinburne Institute of Technology. Faculty of Business

Last Five :

1. Swinburne University of Technology\nFaculty of Life and Social Sciences.\nInstitute for Social Research
2. Swinburne University of Technology\nFaculty of Life and Social Sciences
3. Swinburne University of



Metadata Set: dublin

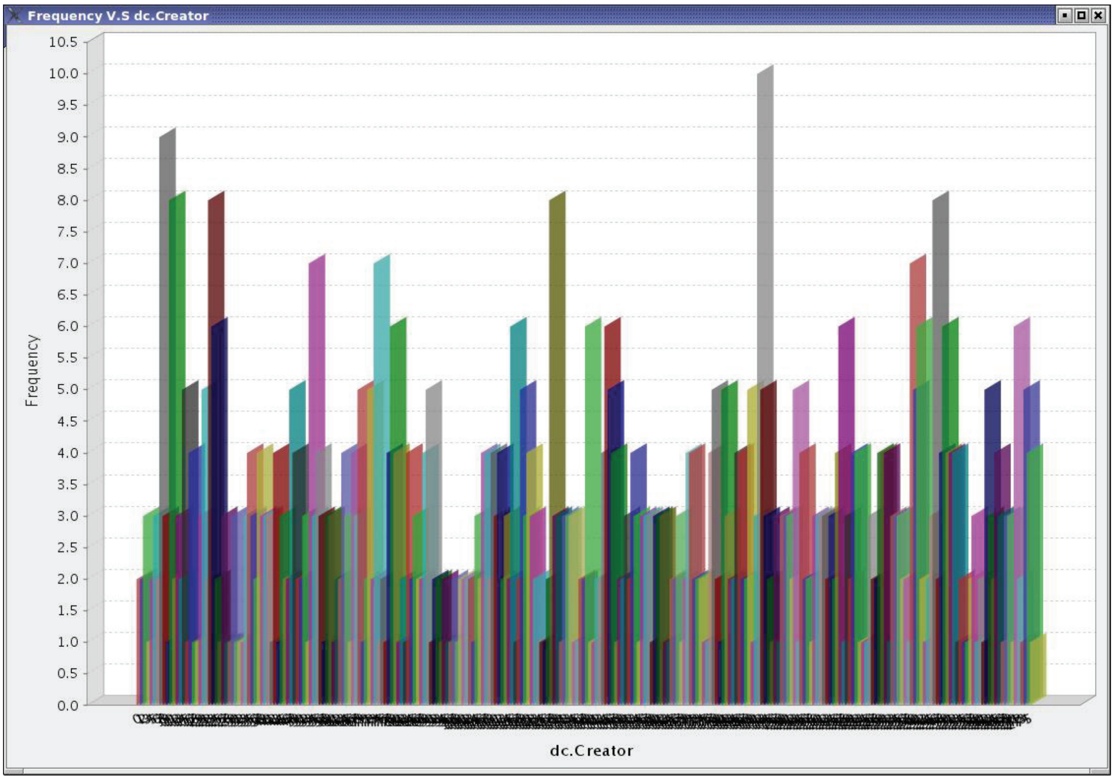
Indexes	Completeness
dc.Title	100.0 %
dc.Creator	100.0 %
dc.Subject	95.6 %
dc.Publisher	94.1 %
dc.Contributor	91.0 %
dc.Date	100.0 %
dc.Type	100.0 %
dc.Format	46.0 %
dc.Identifier	100.0 %
dc.Source	95.4 %
dc.Language	93.1 %
dc.Relation	70.9 %
dc.Coverage	0.0 %
dc.Rights	49.6 %

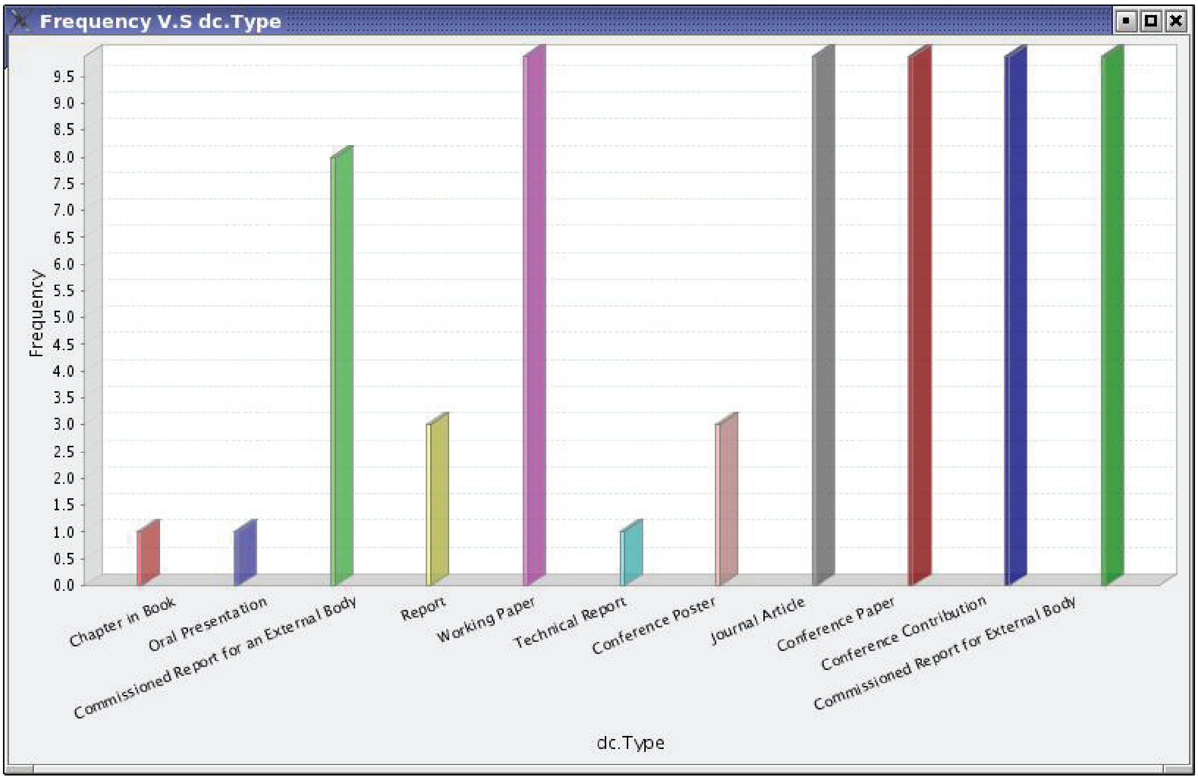
□ Hide Empty Metadata Element

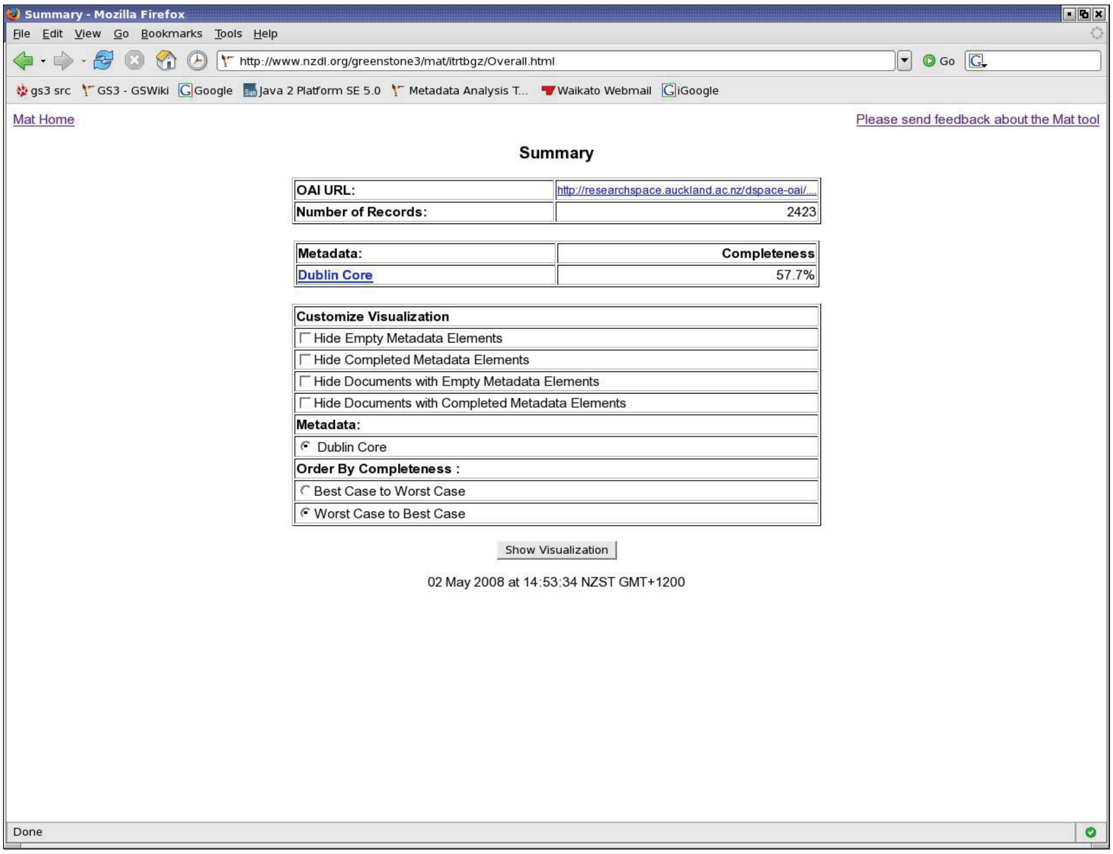
□ Hide Completed Metadata Element

□ Hide Document with empty metadata element...

□ Hide Document with completed metadata ele...


Frequency V.S dc.Creator
Frequency
10.5
10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
dc.Creator


Frequency V.S dc.Type
Frequency
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Chapter in Book
Oral Presentation
Commissioned Report for an External Body
Report
Working Paper
Technical Report
Conference Poster
Journal Article
Conference Paper
Conference Contribution
Commissioned Report for External Body
dc.Type



http://www.nzdl.org/greenstone3/mat/itrtbgz/Overall.html

Summary

OAI URL:	http://researchspace.auckland.ac.nz/dspace-oai/...
Number of Records:	2423

Metadata:	**Completeness**
Dublin Core	57.7%

Customize Visualization

□ Hide Empty Metadata Elements

□ Hide Completed Metadata Elements

□ Hide Documents with Empty Metadata Elements

□ Hide Documents with Completed Metadata Elements

Metadata:

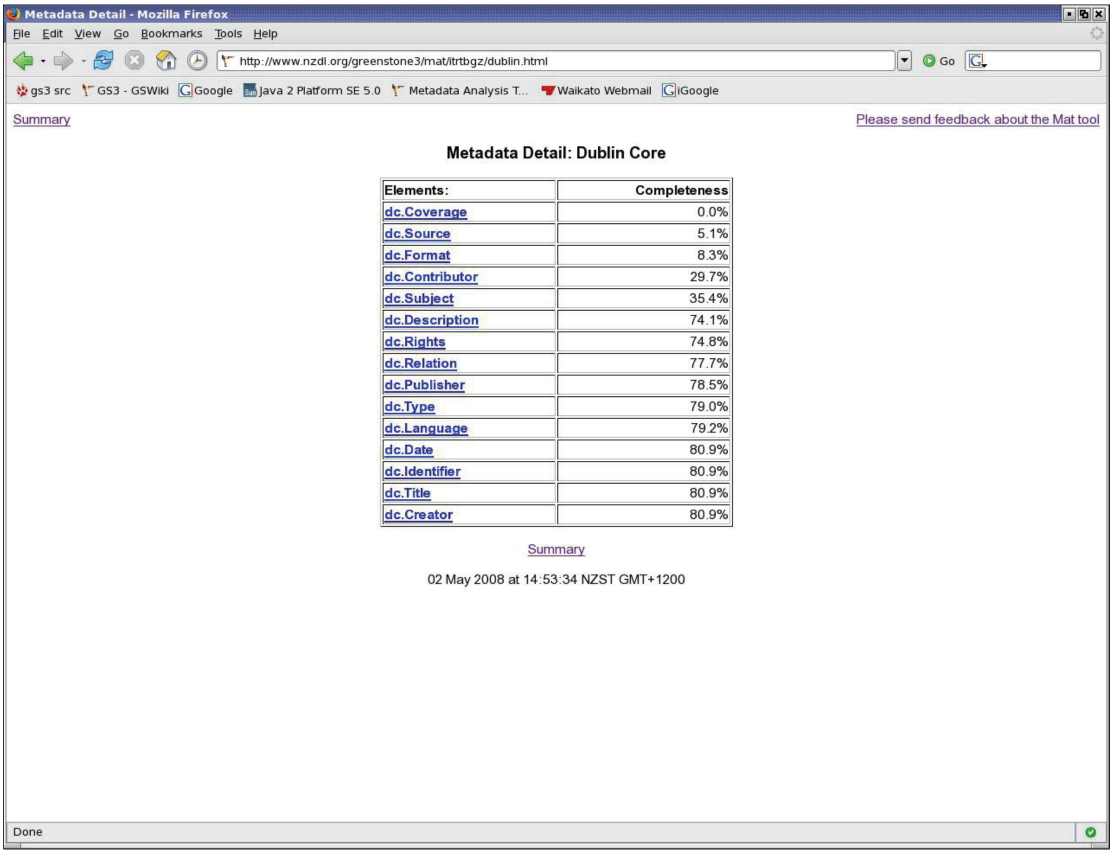
● Dublin Core

Order By Completeness :

○ Best Case to Worst Case

● Worst Case to Best Case

02 May 2008 at 14:53:34 NZST GMT+1200


Metadata Detail - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
gs3 src
GS3 - GSWiki
Google
Java 2 Platform SE 5.0
Metadata Analysis T...
Waikato Webmail
iGoogle
Summary
Please send feedback about the Mat tool
Done

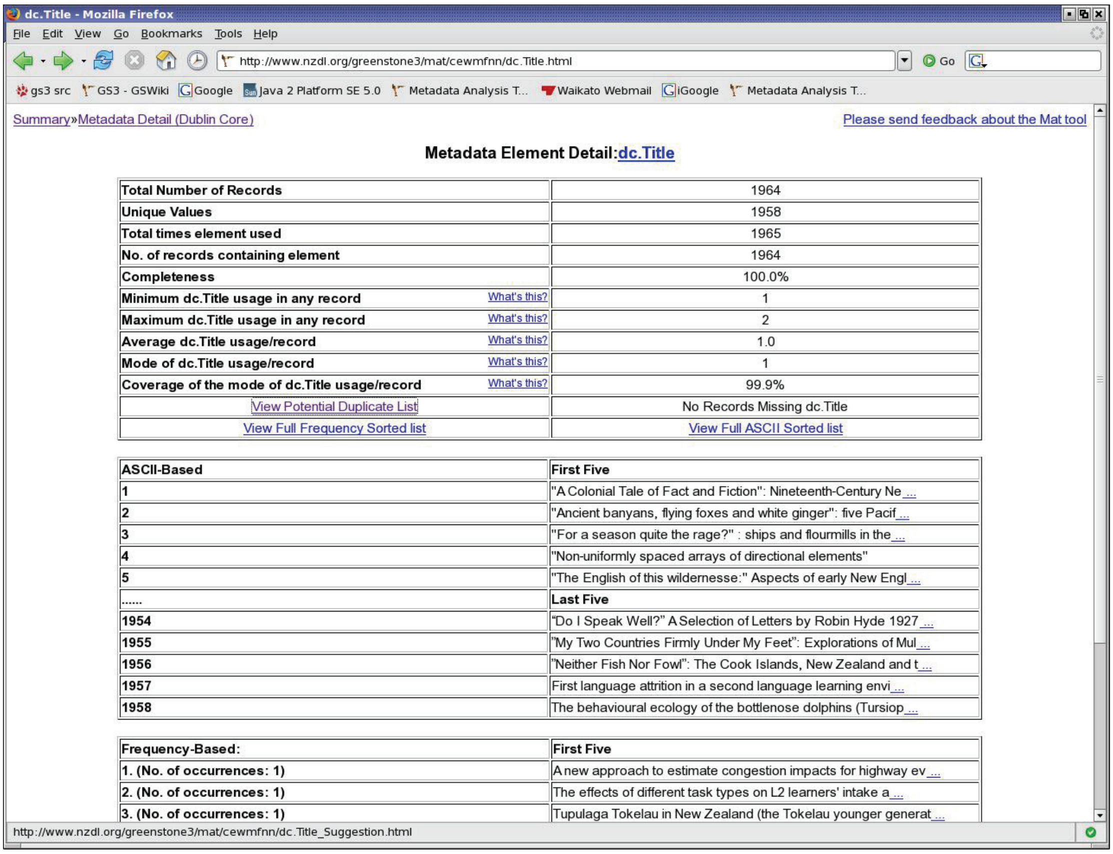
http://www.nzdl.org/greenstone3/mat/itrtbgz/dublin.html

Metadata Detail: Dublin Core

Elements:	**Completeness**
dc.Coverage	0.0%
dc.Source	5.1%
dc.Format	8.3%
dc.Contributor	29.7%
dc.Subject	35.4%
dc.Description	74.1%
dc.Rights	74.8%
dc.Relation	77.7%
dc.Publisher	78.5%
dc.Type	79.0%
dc.Language	79.2%
dc.Date	80.9%
dc.Identifier	80.9%
dc.Title	80.9%
dc.Creator	80.9%

Summary

02 May 2008 at 14:53:34 NZST GMT+1200


dc.Title - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.nzdl.org/greenstone3/mat/cewmfnn/dc.Title.html
Go
gs3 src GS3 - GSWiki Google Java 2 Platform SE 5.0 Metadata Analysis T... Waikato Webmail iGoogle Metadata Analysis T...

Summary»Metadata Detail (Dublin Core)

Please send feedback about the Mat tool

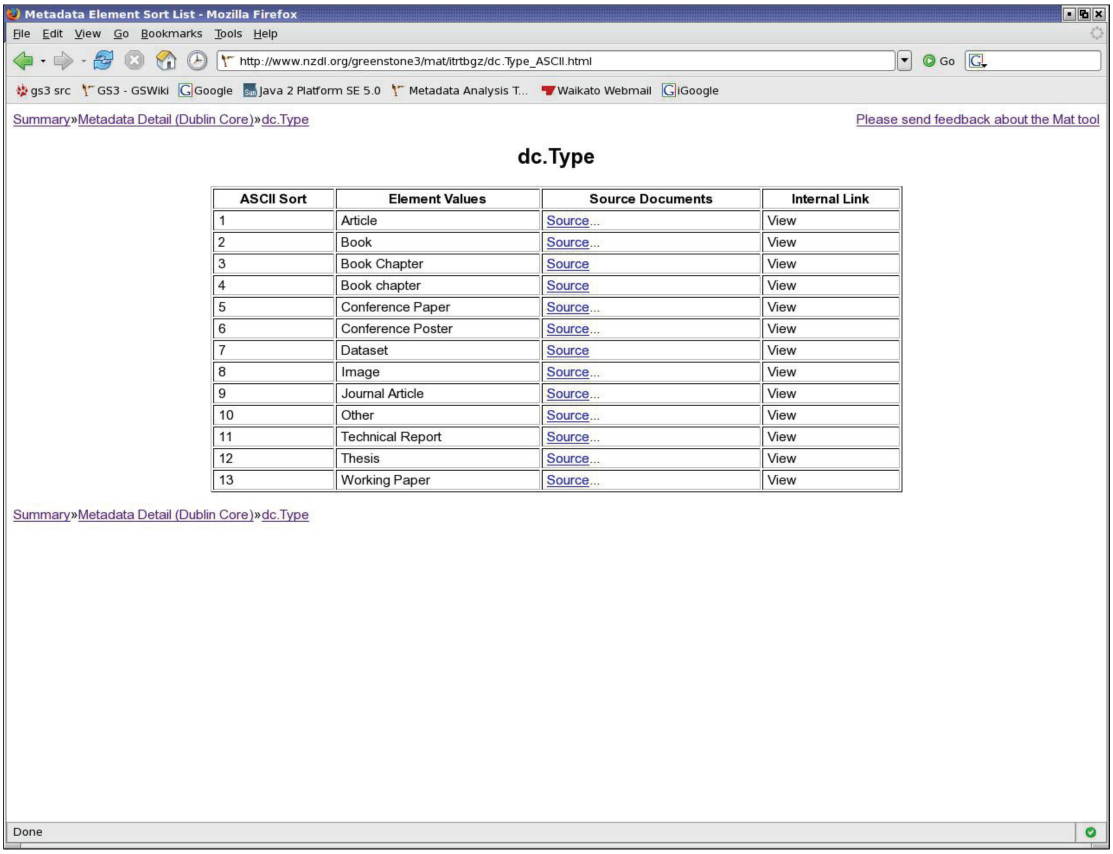
Metadata Element Detail:dc.Title

Total Number of Records	1964
Unique Values	1958
Total times element used	1965
No. of records containing element	1964
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?	99.9%
View Potential Duplicate List	No Records Missing dc.Title
View Full Frequency Sorted list	View Full ASCII Sorted list

ASCII-Based	First Five
1	"A Colonial Tale of Fact and Fiction": Nineteenth-Century Ne ...
2	"Ancient banyans, flying foxes and white ginger": five Pacif ...
3	"For a season quite the rage?" : ships and flourmills in the ...
4	"Non-uniformly spaced arrays of directional elements"
5	"The English of this wildernesse:" Aspects of early New Engl ...
......	Last Five
1954	"Do I Speak Well?" A Selection of Letters by Robin Hyde 1927 ...
1955	"My Two Countries Firmly Under My Feet": Explorations of Mul ...
1956	"Neither Fish Nor Fowl": The Cook Islands, New Zealand and t ...
1957	First language attrition in a second language learning envi ...
1958	The behavioural ecology of the bottlenose dolphins (Tursiop ...

Frequency-Based:	First Five
1. (No. of occurrences: 1)	A new approach to estimate congestion impacts for highway ev ...
2. (No. of occurrences: 1)	The effects of different task types on L2 learners' intake a ...
3. (No. of occurrences: 1)	Tupulaga Tokelau in New Zealand (the Tokelau younger generat ...

http://www.nzdl.org/greenstone3/mat/cewmfnn/dc.Title_Suggestion.html


Metadata Element Sort List - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.nzdl.org/greenstone3/mat/itrtbgz/dc.Type_ASCII.html
Go
gs3 src GS3 - GSWiki Google Java 2 Platform SE 5.0 Metadata Analysis T... Waikato Webmail iGoogle

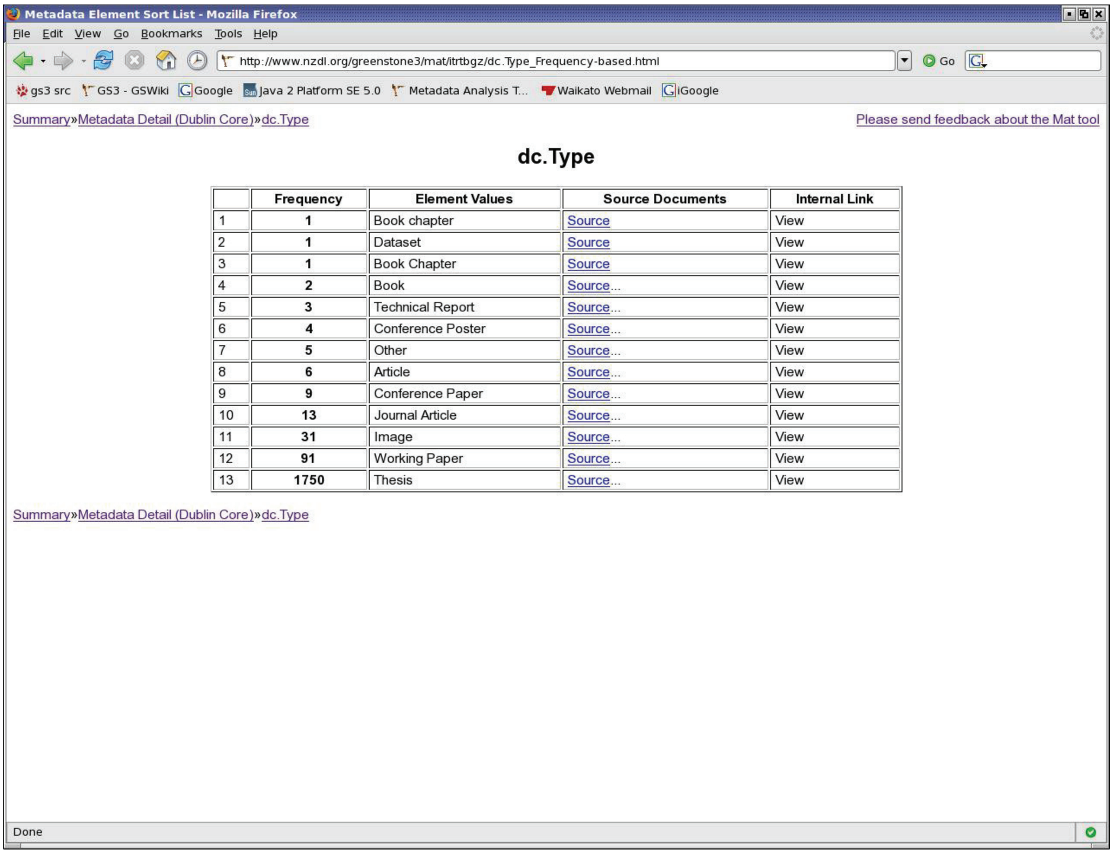
Summary»Metadata Detail (Dublin Core)»dc.Type

Please send feedback about the Mat tool

dc.Type

ASCII Sort	Element Values	Source Documents	Internal Link
1	Article	Source...	View
2	Book	Source...	View
3	Book Chapter	Source	View
4	Book chapter	Source	View
5	Conference Paper	Source...	View
6	Conference Poster	Source...	View
7	Dataset	Source	View
8	Image	Source...	View
9	Journal Article	Source...	View
10	Other	Source...	View
11	Technical Report	Source...	View
12	Thesis	Source...	View
13	Working Paper	Source...	View

Summary»Metadata Detail (Dublin Core)»dc.Type


Metadata Element Sort List - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.nzdl.org/greenstone3/mat/itrtbgz/dc.Type_Frequency-based.html
Go
gs3 src GS3 - GSWiki Google Java 2 Platform SE 5.0 Metadata Analysis T... Waikato Webmail iGoogle
Done

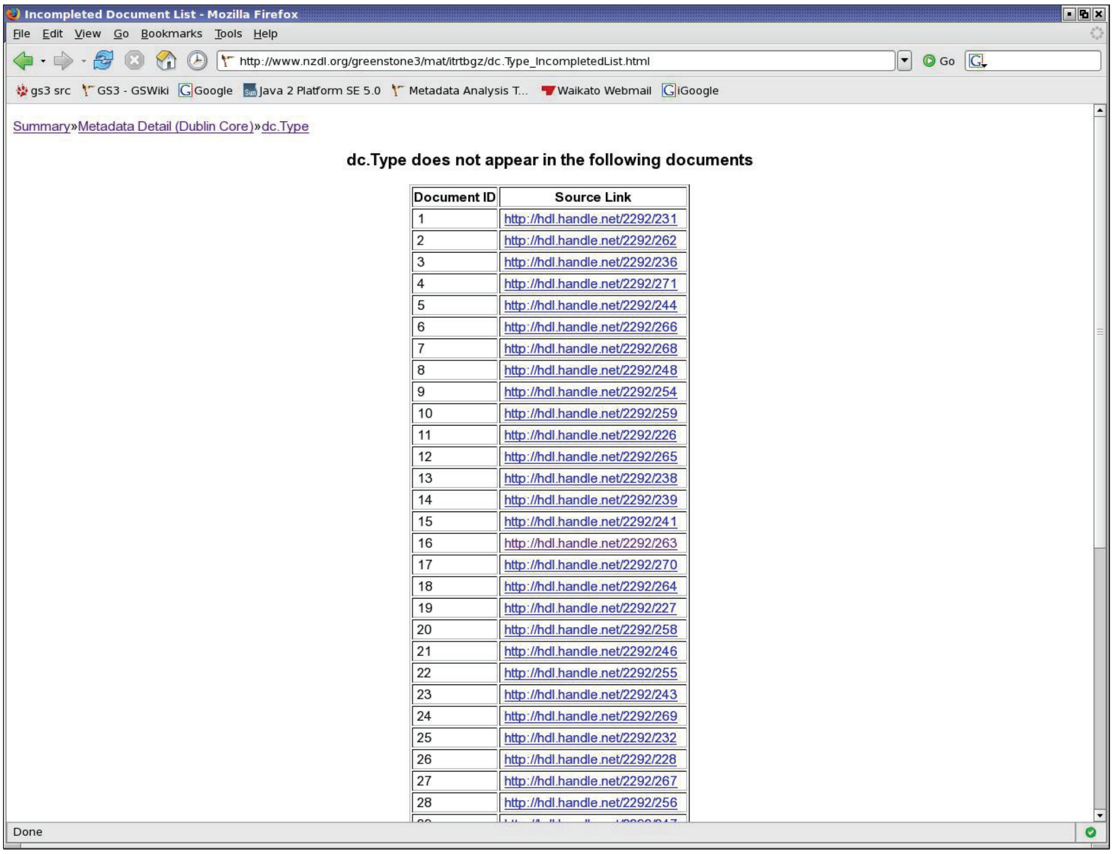
Summary»Metadata Detail (Dublin Core)»dc.Type

Please send feedback about the Mat tool

dc.Type

	Frequency	Element Values	Source Documents	Internal Link
1	**1**	Book chapter	Source	View
2	**1**	Dataset	Source	View
3	**1**	Book Chapter	Source	View
4	**2**	Book	Source...	View
5	**3**	Technical Report	Source...	View
6	**4**	Conference Poster	Source...	View
7	**5**	Other	Source...	View
8	**6**	Article	Source...	View
9	**9**	Conference Paper	Source...	View
10	**13**	Journal Article	Source...	View
11	**31**	Image	Source...	View
12	**91**	Working Paper	Source...	View
13	**1750**	Thesis	Source...	View

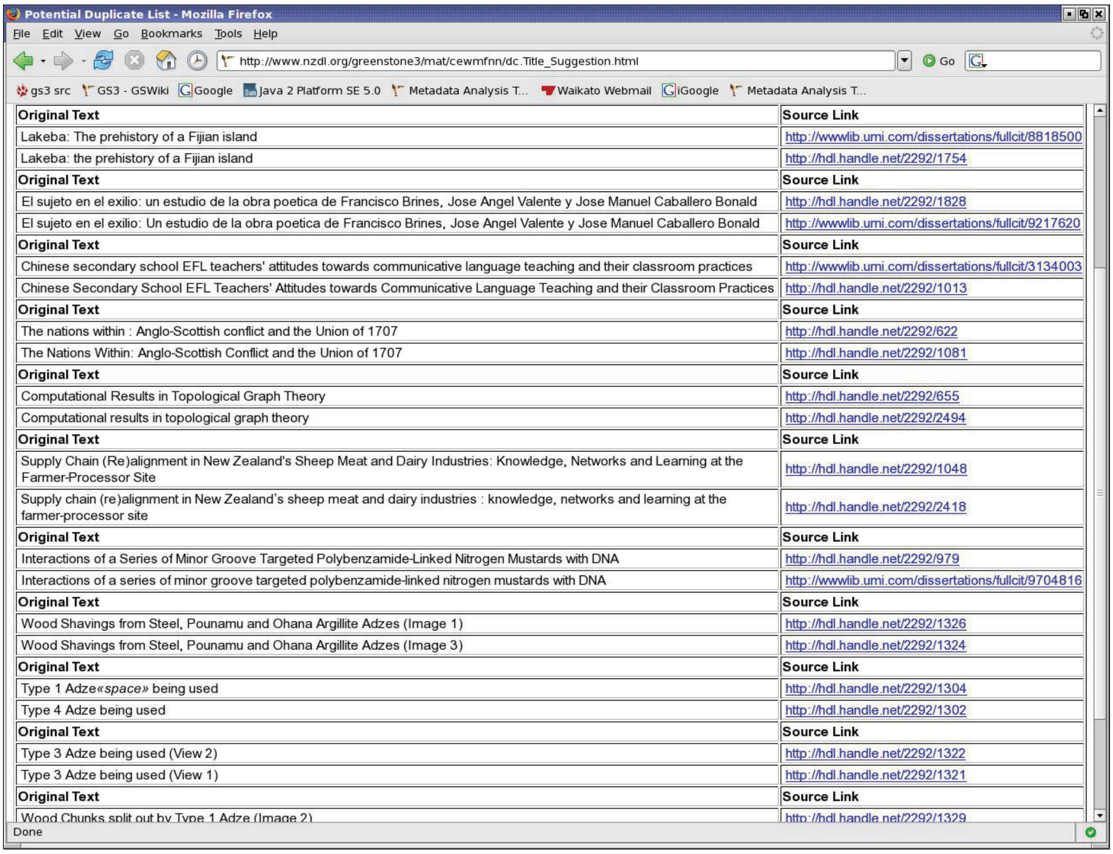
Summary»Metadata Detail (Dublin Core)»dc.Type


Incompleted Document List - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.nzdl.org/greenstone3/mat/itrtbgz/dc.Type_IncompletedList.html
Go
gs3 src GS3 - GSWiki Google Java 2 Platform SE 5.0 Metadata Analysis T... Waikato Webmail iGoogle
Done

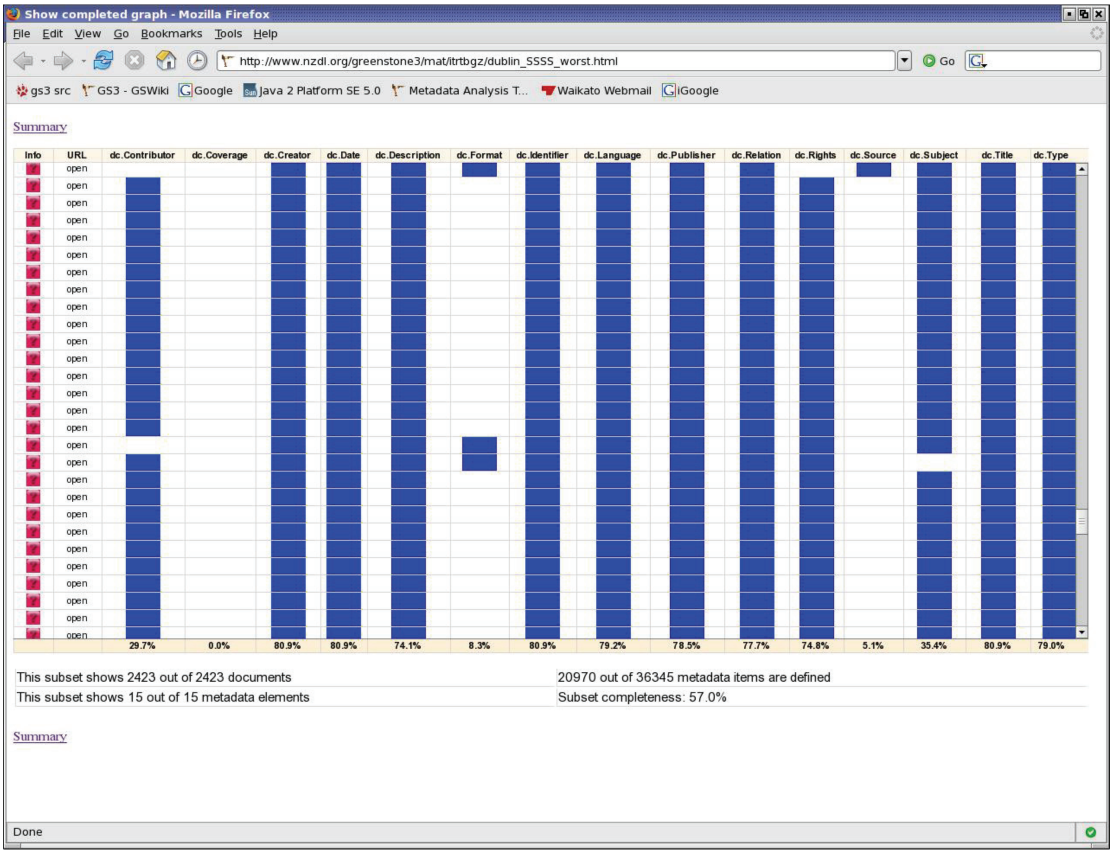
Summary»Metadata Detail (Dublin Core)»dc.Type

dc.Type does not appear in the following documents

Document ID	Source Link
1	http://hdl.handle.net/2292/231
2	http://hdl.handle.net/2292/262
3	http://hdl.handle.net/2292/236
4	http://hdl.handle.net/2292/271
5	http://hdl.handle.net/2292/244
6	http://hdl.handle.net/2292/266
7	http://hdl.handle.net/2292/268
8	http://hdl.handle.net/2292/248
9	http://hdl.handle.net/2292/254
10	http://hdl.handle.net/2292/259
11	http://hdl.handle.net/2292/226
12	http://hdl.handle.net/2292/265
13	http://hdl.handle.net/2292/238
14	http://hdl.handle.net/2292/239
15	http://hdl.handle.net/2292/241
16	http://hdl.handle.net/2292/263
17	http://hdl.handle.net/2292/270
18	http://hdl.handle.net/2292/264
19	http://hdl.handle.net/2292/227
20	http://hdl.handle.net/2292/258
21	http://hdl.handle.net/2292/246
22	http://hdl.handle.net/2292/255
23	http://hdl.handle.net/2292/243
24	http://hdl.handle.net/2292/269
25	http://hdl.handle.net/2292/232
26	http://hdl.handle.net/2292/228
27	http://hdl.handle.net/2292/267
28	http://hdl.handle.net/2292/256


Potential Duplicate List - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.nzdl.org/greenstone3/mat/cewmfnn/dc.Title_Suggestion.html
Go
gs3 src GS3 - GSWiki Google Java 2 Platform SE 5.0 Metadata Analysis T... Waikato Webmail iGoogle Metadata Analysis T...
Done

Original Text	Source Link
Lakeba: The prehistory of a Fijian island	http://wwwlib.umi.com/dissertations/fullcit/8818500
Lakeba: the prehistory of a Fijian island	http://hdl.handle.net/2292/1754
Original Text	**Source Link**
El sujeto en el exilio: un estudio de la obra poetica de Francisco Brines, Jose Angel Valente y Jose Manuel Caballero Bonald	http://hdl.handle.net/2292/1828
El sujeto en el exilio: Un estudio de la obra poetica de Francisco Brines, Jose Angel Valente y Jose Manuel Caballero Bonald	http://wwwlib.umi.com/dissertations/fullcit/9217620
Original Text	**Source Link**
Chinese secondary school EFL teachers' attitudes towards communicative language teaching and their classroom practices	http://wwwlib.umi.com/dissertations/fullcit/3134003
Chinese Secondary School EFL Teachers' Attitudes towards Communicative Language Teaching and their Classroom Practices	http://hdl.handle.net/2292/1013
Original Text	**Source Link**
The nations within : Anglo-Scottish conflict and the Union of 1707	http://hdl.handle.net/2292/622
The Nations Within: Anglo-Scottish Conflict and the Union of 1707	http://hdl.handle.net/2292/1081
Original Text	**Source Link**
Computational Results in Topological Graph Theory	http://hdl.handle.net/2292/655
Computational results in topological graph theory	http://hdl.handle.net/2292/2494
Original Text	**Source Link**
Supply Chain (Re)alignment in New Zealand's Sheep Meat and Dairy Industries: Knowledge, Networks and Learning at the Farmer-Processor Site	http://hdl.handle.net/2292/1048
Supply chain (re)alignment in New Zealand's sheep meat and dairy industries : knowledge, networks and learning at the farmer-processor site	http://hdl.handle.net/2292/2418
Original Text	**Source Link**
Interactions of a Series of Minor Groove Targeted Polybenzamide-Linked Nitrogen Mustards with DNA	http://hdl.handle.net/2292/979
Interactions of a series of minor groove targeted polybenzamide-linked nitrogen mustards with DNA	http://wwwlib.umi.com/dissertations/fullcit/9704816
Original Text	**Source Link**
Wood Shavings from Steel, Pounamu and Ohana Argillite Adzes (Image 1)	http://hdl.handle.net/2292/1326
Wood Shavings from Steel, Pounamu and Ohana Argillite Adzes (Image 3)	http://hdl.handle.net/2292/1324
Original Text	**Source Link**
Type 1 Adze«*space*» being used	http://hdl.handle.net/2292/1304
Type 4 Adze being used	http://hdl.handle.net/2292/1302
Original Text	**Source Link**
Type 3 Adze being used (View 2)	http://hdl.handle.net/2292/1322
Type 3 Adze being used (View 1)	http://hdl.handle.net/2292/1321
Original Text	**Source Link**
Wood Chunks split out by Type 1 Adze (Image 2)	http://hdl.handle.net/2292/1329



This subset shows 2423 out of 2423 documents	20970 out of 36345 metadata items are defined
This subset shows 15 out of 15 metadata elements	Subset completeness: 57.0%

Appendix C: Survey URLs

This appendix contains the URLs for repository survey in section 5.3.

Collection 1

Repository Name: ResearchBank

URL: http://researchbank.swinburne.edu.au:8080/fedora/oai

Collection 2

Repository Name: MINDS @ UW

URL: http://minds.wisconsin.edu/oai/request

Collection 3

Repository Name: ResearchSpace at The University of Auckland

URL: http://researchspace.auckland.ac.nz/dspace-oai/request

Collection 4

Repository Name: Otago Eprints

URL: http://eprints.otago.ac.nz/perl/oai2

Collection 5

Repository Name: IDEALS @ UIUC

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

Collection 6

Repository Name: Research Commons

URL: http://waikato.researchgateway.ac.nz/dspace-oai/request