Working Paper Series
ISSN 1177-777X

A TOOL FOR METADATA ANALYSIS

David M. Nichols
Chu-Hsiang Chan
David Bainbridge
Dana McKay
and
Michael B. Twidale

Working Paper: 02/2008
February 2008

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

A Tool for Metadata Analysis

David M. Nichols,
Chu-Hsiang Chan, David Bainbridge
Department of Computer Science
University of Waikato
Hamilton, New Zealand
+64 (7) 8585130
{dmn,cc108,davidb}@cs.waikato.ac.nz

Dana McKay
University Library
Swinburne University of Technology
Hawthorn, VIC 3122
Australia
+61 (3) 9214 5023
dmckay@swin.edu.au

Michael B. Twidale
Graduate School of Library and Information Science
University of Illinois
Champaign, IL 61820, USA
+1 (217) 265-0510
twidale@uiuc.edu

ABSTRACT
We describe a Web-based metadata quality tool that provides statistical descriptions and visualisations of Dublin Core metadata harvested via the OAI protocol. The lightweight nature of development allows it to be used to gather contextualized requirements and some initial user feedback is discussed.

Categories and Subject Descriptors
D.3.3 [**Information Storage and Retrieval**]: Digital Libraries – *user issues, systems issues.*

General Terms
Measurement, Human Factors.

Keywords
Metadata quality, OAI, visualisation, Dublin Core, Prototyping

1. INTRODUCTION
The growth in the number, size and diversity of digital collections makes metadata quality an increasingly important issue. Consequently, appropriate software tools offer great potential for collection managers to analyse their repositories and verify that their metadata supports their users' interactions.

In this paper we describe a lightweight metadata quality tool for OAI repositories that enables users to analyse their collections. Although the motivation for the tool is to help gather requirements for future functionality we have found that in its prototype form it is already useful for detecting metadata quality issues [2].

Section 2 of the paper gives some background on metadata quality and existing visualization tools. Section 3 describes the tool, followed by some initial user feedback.

2. BACKGROUND
Bruce and Hillmann [3], whilst acknowledging the difficulties in defining metadata quality, list seven metadata quality criteria: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. Assessing these criteria in diverse collections requires both human expertise and appropriate tool support [2, 8]. The potential value of metadata quality tools for collection managers is clear: "automated techniques potentially enable humans to use their time to make more sophisticated assessments" [3]. Some of these quality criteria are particularly amenable to computational evaluation, such as completeness, whereas others, such as timeliness, are better suited to human judgment [2, 3].

An important category of supportive tools are those that produce visualisations: graphic depictions of data that allow human visual processing to quickly make complex judgments: "the use of data visualization software can significantly improve efficiency and thoroughness of metadata evaluation" [4]. Despite the enthusiasm and promise of the Dushay & Hillmann paper in 2003 [4] there appears to be little evidence that repository managers are using visualisation or quality analysis tools to improve their collections. One notable exception is work at UDLA where starfield displays [1] have been used to visualize library catalogue data [11]. Recently, a treemap visualisation has been used for locating areas of a collection with lower metadata quality [8]. However, these systems have not been widely deployed and so there are no authentic reports from repository managers in the literature. Recent references to visual tools appear to be in the form of sample use case rather than actual experience reports [7]. There are several possible reasons for this lack of adoption including: no awareness of the problem or of the possible tools, lack of money [10] to purchase information visualization tools, complexity of integration of the tool with the data set, lack of time to learn a complex application [10] etc.

Several surveys (e.g. [12, 15]) appear to use custom-written software to analyse OAI repositories, focusing on sampling and/or processing to give overall summary information rather than visualisations. Shreeves *et al.* [12] looked at the impact of OAI federation processes on the nature of overall metadata quality. With federated collections, it was found that the act of federating itself can degrade metadata quality for various reasons. Firstly, multiple metadata formats may need to be combined and reconciled into the single format of the union collection. Secondly, different projects even if using the same format, will evolve their own norms for usage that mean that although their collection is consistent in that usage, the combined collection will have inconsistencies that can confuse users and render search actions problematic. Thirdly, a given collection may have assumed knowledge given its nature and location that, precisely because it is assumed, is not explicitly represented in the metadata. This need not cause an access problem in its local contextualized use, but does cause problems in federated use.

In summary, there is significant potential for metadata quality tools to allow collection managers to improve their repositories [13]. For a variety of reasons tools for quality analysis appear not to be widely deployed or used. However, as OAI harvesting projects have shown, the technical barriers to providing a public

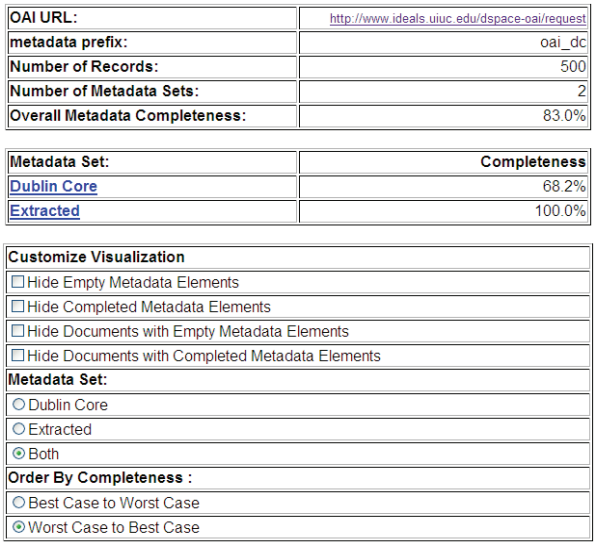

Figure 1. The summary view of 500 OAI metadata records from IDEALS @ Illinois

tool are not large. In the next section we describe how we leveraged existing Greenstone functionality to provide an online metadata quality analysis service.

3. TOOL DESCRIPTION

The original goal of this project was to provide a quality analysis component that could become part of the Greenstone Librarian Interface (GLI) [16]. However, we found few experience reports in the literature and so shifted focus to better understand the needs of collection managers. We chose to build and deploy a prototype tool as the most effective mechanism to solicit user feedback as we agree that "[metadata] tool development needs to be an iterative process between developers and users" [6].

Although GLI is a Java application we chose a Web deployment to reduce technology barriers to use [5] so that we could in turn gather software requirements from a *wide* group of potential adopters (beyond current Greenstone users). Additionally, by providing a free service we allow repository managers to use their own data and so avoid some of the problems of earlier evaluation approaches: "usability of information visualization tools can be measured in a laboratory however, to be convincing, utility needs to be demonstrated in a real settings … Using real datasets with more than a few items, and demonstrating realistic tasks is important" [9]. Thus, the prototype supports rapid, incremental requirements capture based on small scale contextualized use.

3.1 Leveraging Greenstone Technology

The tool is constructed on top of Greenstone3, using a servlet in Apache Tomcat. The servlet communicates with existing Greenstone command line tools for collection building and then outputs static HTML quality evaluation reports. The core statistical code of the tool is common to the Java application and the Web version. Our deployment approach is similar to the complementary service of the OAI Repository Explorer [14], our tool is available at:

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

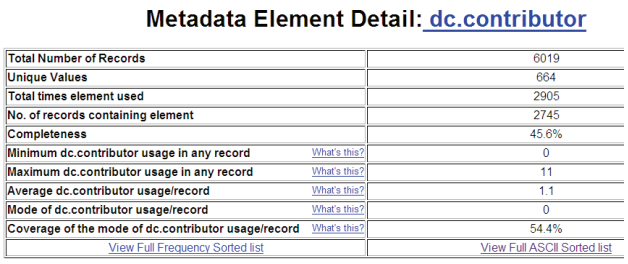

Figure 2. Part of the element detail view

3.2 Tool Features

The tool has three main features intended to aid collection managers: summary description of metadata elements, sorted presentation of metadata element lists and a completeness-oriented visualisation. Figure 1 shows the initial page of a completed report, including some summary statistics and the customisation options for the visualisation. If the user selects the 'Dublin Core' link then a completeness-oriented summary is shown, with entries such as:

```
dc.format.medium              4%
dc.coverage.temporal        100%
```

The user can then 'drill down' to the specifics of a particular element (Figure 2). This element detail view provides some simple descriptive statistical measures for the element, including its average occurrence. This view also shows a sampling of the frequency and ASCII sorting (not shown in Figure 2), full versions are presented on separate pages. The choices of ASCII and frequency ordering were heuristics we though that collection managers would find useful and we expect different types of sorting to be developed as the tool evolves.

On the left of Figure 4 are the most frequent values of a `dc.language` element, which might suggest some inconsistency for documents in English. On the right of Figure 4 are the last few values of an ASCII sort on a `dc.type` element, showing spurious newline characters 'sinking' to the bottom of an ASCII ordering (leading space characters 'float' to the top).

The visualisation element of our online tool (Figure 3) closely resembles the example scatter plot of metadata from the Spotfire application described in [4]. Focusing on subsets of the data is an important aspect of metadata visualisations and advanced tools such as Spotfire have several mechanisms for customising their displays. The initial tool deployment has some simple options to whet users' appetites and encourage useful feedback (Figure 1): such as sorting documents by metadata completeness and hiding metadata elements that are complete (or empty). These options reduce the number of data points displayed, with two main benefits: smaller displays are much easier to manage in the constrained environment of a web browser and they allow users to focus on partially complete records/elements. These options attempt to support the requirement that "an evaluator could easily focus on where it was useful to look more closely at the data" [4].

Figure 3 shows 13 Dublin Core elements (as two empty elements have been hidden) and 6000 records in a scrolling table. The presence of a metadata item is indicated by a blue rectangle with white areas indicating undefined metadata items. On the left of the visualisation is a button to show the full metadata for a record and a link (heuristically extracted from `dc.identifier`) back to

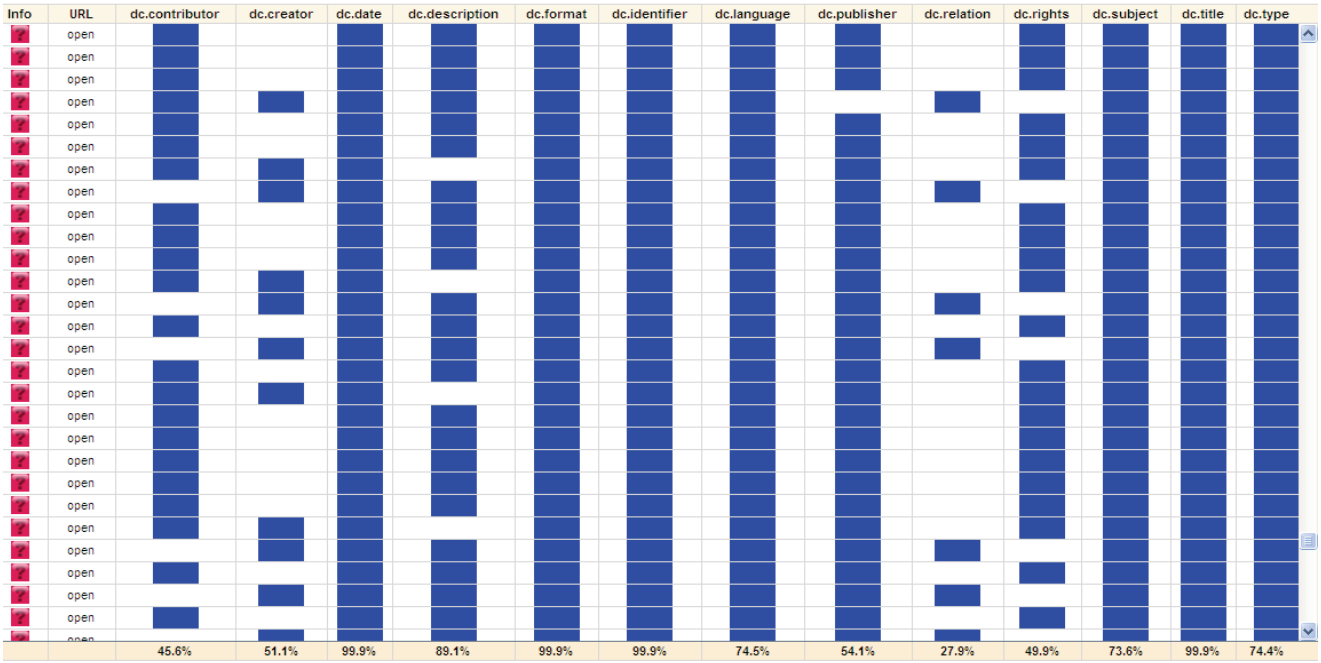

Figure 3. Part of a visualisation of 6000 OAI Dublin Core records from MINDS @Wisconsin (two empty elements are hidden)

1	fr
4	es
46	ur
350	en
1181	N/A
1616	English
2902	en_US

Thesis (Honours)
Thesis (MBA Project)
Thesis (PhD)
Working paper
\n Technical report\n
\nBook chapter\n
\nConference paper\r
\nJournal article\n

Figure 4. Excerpts from an element frequency sort (left) and an ASCII sort (right)

the item in the remote repository. The records in Figure 3 have been sorted by completeness; with the records missing more metadata at the top; it is thus a specific example of the suggested "visual view" approach to metadata quality [4]. The 6000 records and 13 elements in Figure 3 require an HTML page of about 2Mb.

The visualization is far less sophisticated than Spotfire, nevertheless, it allows for interesting use. Rapidly scrolling a page gives an overall sense of the 'shape' of the quality landscape. For example: does it seem that the quality is fairly uniform? Does it degrade smoothly? Is there a rump of very poor records or possibly null records? These are crude indicators, but they can serve as a starting point for detailed investigation. Without scrolling, a more focused comparison of a few individual records becomes possible. Does coverage strongly correlate between all fields or certain fields? Are some fields in 'antiphase'?

There are several possible relationships in the visualisation example of Figure 3, for example `dc.rights` and `dc.contributor` tend to be empty when `dc.creator` and `dc.relation` are defined. This could represent two different standards for metadata entry from two different collections or two different individuals – there is the possibility that the antiphase relation is caused by a consistent disagreement about which field to use to enter a particular value. It might be due to local policy for different types of records or it might be an indicator that metadata entry is not following agreed standards. A combination of domain knowledge and more advanced sorting/grouping methods are needed to investigate further.

4. RESULTS AND DISCUSSION

The tool typically reports two metadata sets in its results: the main Dublin Core and an additional extracted set. This automatically extracted metadata is generally not that valuable as it is derived from internal Greenstone processing of downloaded OAI files. Completeness is usually 100% on these automatically assigned terms; however the tool's reports alerted the developers to a specific case where titles were not assigned. This was a useful result even before the tool had been deployed.

4.1 Prototype results and feedback

We gathered initial feedback about the tool using an online survey, and semi-structured interviews with repository managers. Most feedback received was from repository managers, though one self-identified software developer contributed a survey response. All but one of the repository managers who contributed feedback were actively managing a repository; the final manager (Participant C) was still planning development of his repository. . Respondents worked variously with DSpace and Fedora-based repositories; some managers' repositories used self-deposit, and some deposited all work themselves.

Participants were generally very excited by the tool, seeing great potential for collection improvement (particularly as their own repositories do not offer similar tools). Despite a long list of

feature requests, feedback was generally positive, and in the words of Survey Respondent A: "it is so nice to see people working in this area. Well done!" The comments can be divided into three categories: metadata quality, potential uses for the tool, and feature requests.

4.1.1 Metadata Quality

All the repository managers interviewed were very concerned with metadata quality, and as one manager commented "metadata completeness is a mark of record quality". They were impressed with the at-a-glance depictions of metadata completeness, and also with the ability to see what kinds of metadata were in their repositories using the list views for individual metadata elements. All repository managers had been involved in metadata translation from another schema at some point (Participant C commented "I really never think in DC, it's only used when you need interoperable metadata") and the tool was seen to be an excellent way of checking the quality of metadata translations.

4.1.2 Potential Uses

Participants and survey respondents who were not actively managing a repository found it considerably more difficult to imagine uses than did active repository managers. Active repository managers all mentioned that it would be valuable to use to check metadata completeness at periodic intervals. Other uses mentioned included checking that an OAI feed was working correctly after a software upgrade, improving metadata entry practices, and generating repository statistics not available from their repository software. 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. This episode serves as a reminder that the tool needs to work in conjunction with repositories rather than in a stand-alone manner.

4.1.3 Feature requests

All respondents had some feature requests; some wanted the tool to deal with different types of metadata (though one said he thought that was a waste of time, given the number of different schemas in use). Some repository managers asked for more documentation (and others missed features during an initial exploration), so usability improvements and documentation are clearly a priority for further development. Many participants wanted links from the sorted element views to the associated documents, so they could immediately repair incorrect metadata, and most of them would like the tool to work faster to build the reports.

5. CONCLUSIONS

We have described a web-based metadata analysis tool that leverages Greenstone3 technology to provide a service for collection managers. The tool shows promise, both as a method for gathering requirements for metadata quality tools and as a supplement to digital library education activities. When students build digital collections the main form of feedback is the built collection itself; a supplementary objective automated quality assessment could be a valuable pedagogical tool for digital library educators. Planned future work involves more user studies, exploring machine-readable application profiles [7] and heuristic error detection using approximate string matching.

6. REFERENCES

[1] Ahlberg, C. and Shneiderman, B. 1994. Visual information seeking: tight coupling of dynamic query filters with starfield displays. *Proceedings of CHI'94.* ACM. 313-317.

[2] Beall, J. 2005. Metadata and data quality problems in the digital library. *Journal of Digital Information* 6(3).

[3] 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.

[4] Dushay, N, and Hillmann, D.I. 2003. Analyzing metadata for effective use and re-use. *Proceedings of the International Conference on Dublin Core and Metadata Applications*, Seattle, WA.

[5] Golub, E. and Shneiderman, B. 2003. Dynamic query visualizations on World Wide Web clients: a DHTML solution for maps and scattergrams. *International Journal of Web Engineering and Technolology,* 1(1) 63-78.

[6] Greenberg, J. and Severiens, T. 2006. Metadata Tools for Digital Resource Repositories: JCDL 2006 Workshop Report. *D-Lib Magazine*, 12(7/8).

[7] Hillmann, D.I. and Phipps, J. 2007. Application profiles: exposing and enforcing metadata quality, *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2007)*, 52-62. Singapore.

[8] Ochoa, X. and Duval, E. 2006. Towards automatic evaluation of learning object metadata quality, In *Advances in Conceptual Modeling - Theory and Practice, ER 2006 Workshops BP-UML, CoMoGIS, COSS, ECDM, OIS, QoIS, SemWAT*, Springer, 372-381.

[9] Plaisant, C. 2004. The challenge of information visualization evaluation. *Proceedings of the Working Conference on Advanced Visual interfaces (AVI '04).* ACM. 109-116.

[10] Salo, D. 2007. Innkeeper at the Roach Motel, (to appear in *Library Trends*) http://digital.library.wisc.edu/1793/22088

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

[12] Shreeves, S., Knutson, E., Stvilia, B., Palmer, C., Twidale, M. and Cole, T. 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.

[13] Stvilia, B., Gasser, L., Twidale M.B. and Smith L.C. 2007. A framework for information quality assessment. *JASIST*, 58(12), 1720-1733.

[14] Suleman, H. 2001. Enforcing interoperability with the open archives initiative repository explorer. *Proceedings of JCDL '01.* ACM. 63-64.

[15] Ward, J. 2004. Unqualified Dublin Core usage in OAI-PMH data providers. *OCLC Systems & Services*, 20(1), 40-47.

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