Working Paper Series
ISSN 1177-777X

METADATA TOOLS FOR INSTITUTIONAL REPOSITORIES

David M. Nichols, Gordon W. Paynter, Chu-Hsiang Chan, David Bainbridge, Dana McKay, Michael B. Twidale and Ann Blandford

Working Paper: 10/2008
August 2008

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand

Metadata Tools for Institutional Repositories

David M. Nichols [1], Gordon W. Paynter [2], Chu-Hsiang Chan [1],
David Bainbridge [1], Dana McKay [3], Michael B. Twidale [4] and Ann Blandford [5]

[1] Department of Computer Science, University of Waikato, Hamilton, New Zealand

[2] National Library of New Zealand, Wellington, New Zealand

[3] University Library, Swinburne University of Technology, Hawthorn, Australia

[4] Graduate School of Library and Information Science, University of Illinois, USA

[5] UCL Interaction Centre, University College London, UK

Abstract

Current institutional repository software provides few tools to help metadata librarians understand and analyse their collections. In this paper we compare and contrast metadata analysis tools that were developed simultaneously, but independently, at two New Zealand institutions during a period of national investment in research repositories: the Metadata Analysis Tool (MAT) at The University of Waikato, and the Kiwi Research Information Service (KRIS) at the National Library of New Zealand.

The tools have many similarities: they are convenient, online, on-demand services that harvest metadata using OAI-PMH, they were developed in response to feedback from repository administrators, and they both help pinpoint specific metadata errors as well as generating summary statistics. They also have significant differences: one is a dedicated tool while the other is part of a wider access tool; one gives a holistic view of the metadata while the other looks for specific problems; one seeks patterns in the data values while the other checks that those values conform to metadata standards.

Both tools work in a complementary manner to existing web-based administration tools. We have observed that discovery and correction of metadata errors can be quickly achieved by switching web browser views from the analysis tool to the repository interface, and back. We summarise the findings from both tools' deployment into a checklist of requirements for metadata analysis tools.

Keywords: metadata quality, institutional repositories

1. Introduction

Current institutional repository software provides few tools for metadata librarians to understand and analyse their collections. In this paper we compare and contrast two metadata analysis tools for repositories that address this lack. Both tools harvest metadata using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and help metadata librarians analyse this data, pinpointing specific metadata errors and generating summary statistics.

The Kiwi Research Information Service (KRIS) is provided by the National Library of New Zealand to help disseminate research outputs from the New Zealand tertiary sector. To help ensure quality the tool validates the harvested metadata against agreed national guidelines and provides periodic and on-demand reports for managers analysing their repository's compliance (http://nzresearch.org.nz/).

The Metadata Analysis Tool (MAT) is built on top of the Greenstone digital library software and provides a public service for analysing OAI collections (http://nzdl.org/greenstone3/mat). Metadata analysis reports are generated that provide an alternative element-centric view of a repository using pre-defined sorting heuristics. A visualisation of the metadata distribution is also provided to support discovery of patterns of anomalies.

In this paper, we describe the issues involved in deploying and maintaining these online tools. Qualitative feedback, through surveys and interviews on the use of the tools, has provided useful feedback for further clarifying the requirements for metadata analysis tools. Repository managers appreciate the alternative external views of their collections provided by tools using harvesting approaches. However, the analysis functionality is constrained by repositories that only make available a 'dumbed-down' subset of their full metadata (i.e. unqualified Dublin Core).

The reports produced by KRIS are valued as managers can refer to up-to-date results at any time, and also support national policymakers by producing an estimate of the "state of the nation's metadata". The features provided by MAT, such as browsable sorted lists of elements, can be surprisingly useful even when the sorting criteria are relatively simple. Sorting by frequency and by ASCII-ordering allows several types of errors to either float to the top or sink to the bottom of result lists; so becoming easier to identify. The visualisation component provides a high-level view of completeness for a repository which complements the element-centric approaches and is a preferred starting point for collection exploration by some managers.

Section 2 gives an outline of the literature on tools to support metadata analysis. We then describe the two analysis tools we have deployed and show examples of their output. In Section 5 we outline our experiences in designing and deploying the systems. We then discuss our findings and conclude with a checklist of requirements for metadata analysis tools.

2. Background

The rapid growth of institutional repositories (IRs) has been facilitated through software tools such as DSpace (Tansley *et al.,* 2005) and EPrints (EPrints, 2008). These tools have lowered entry barriers for organisations wishing to make resources accessible via the Web. However, in practice the repository managers are often marginalised within libraries, are left without

sufficient technical support and have to deal with poorly designed software tools (Salo, 2008 to appear). If we accept that "supporting the development of quality metadata is one of the most important roles for LIS professionals" (Robertson, 2005) then the available tools are constraining the ability of library staff to adapt their skills to the new setting of IRs.

All activities of metadata creation need to consider issues of quality, data checking, error correction and the ongoing refinement of processes for error prevention, but in the case of IRs there can be circumstances where tradeoffs are consciously made to lower quality (temporarily) in order to achieve other valuable features such as coverage. There are a number of challenges in setting up an IR. To be useful it typically needs to be both easily accessible through accurate and substantial metadata, but also to have a reasonably good coverage of the collection. In the absence of the former, users will fail to find what is actually in the collection, but in the absence of the *perception* of good coverage, users may not even bother searching the collection in the first place. One approach to the challenge of coverage is to aggregate or federate collections, even though this is known to have a somewhat negative effect on data quality (Shreeves *et al.,* 2005). Another approach may be a willingness to accept a somewhat lower than ideal initial level of data quality in order to enable rapid early growth of the IR, encouraging its visibility, and enabling the cultural change necessary to the adoption of the new activities needed to maintain the IR. Inevitably newcomers will make errors in creating metadata, and if the creation of the metadata is partially or wholly in the hands of content creators rather than professional cataloguers, the error rate will be higher still. Over time, these initial errors can be corrected and participants can learn how to improve the quality of metadata at the point of creation. It is a matter for repository managers to decide the extent to which they want their repository to be more like a traditional collection catalogue (very accurate, but often slow to appear) or more like Wikipedia or beta release software (very rapid and responsive, but acknowledged to have a substantial number of errors). Quality visualisation tools are useful whichever point on the quality-speed continuum an IR is positioned.

Beall (2005) surveys quality issues for both data and metadata in digital collections, re-iterating that poor quality metadata impedes access to resources. The article also provides a discussion of the types of metadata error, the responsibility for errors, strategies for handling errors and outlines various practices through which errors can be introduced. However, there is an absence of discussion of the *discovery* of metadata errors but an implicit recognition that the size of digital collections means that manual techniques will be infeasible. Bruce and Hillmann (2004) are explicit: "automated techniques potentially enable humans to use their time to make more sophisticated assessments [of metadata quality]".

Bruce and Hillmann (2004) list seven metadata quality criteria: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. Criteria such "conformance to expectations" clearly require human judgement whereas others such as "completeness" are amenable to computational evaluation. There is some evidence that relatively easy to compute measures such as completeness correlate reasonably well with the more useful but more complex measures of quality (Stvilia *et. al*, 2007), at least hinting at areas of the dataset that might be more problematic and would repay more detailed examination. Furthermore, certain absences and errors have a much greater impact on the findability of records than others. The details vary from collection to collection, and so again rely on informed judgement to decide which errors and omissions it is most cost effective to remedy. In practice then, neither a manual nor an automated approach alone is sufficient and we should aim for supportive tools that empower repository managers to effectively assess and address issues of metadata quality in their collections.

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" (Dushay and Hillmann, 2003). Despite the enthusiasm and promise of the Dushay and Hillmann paper there appears to be little evidence that repository managers are using visualisation or quality analysis tools to investigate their collections.

Although repository managers seem not to be using automated tools to inspect their local collections, several surveys have used the OAI-PMH to investigate metadata in remote repositories. Dublin Core element usage data has been compared over many repositories (Efron 2007; Shreeves *et al.* 2005; Ward 2004) but it appears that these surveys have used custom-written software. Additionally these approaches had the aim of analysing element usage which, although similar, is not the same task as metadata analysis oriented towards quality through detection and correction of records containing errors.

A further distinction between different OAI-PMH tools can be made between those that analyse the implementation of the protocol versus those that examine the values of the content retrieved via the protocol. In this paper (as with the Dublin Core usage surveys) we are concerned with content and do not address issues of protocol validation, which are best dealt with by specific tools such as the *OAI Repository Explorer* (Suleman, 2001).

In summary, there is significant potential for metadata quality tools to allow collection managers to improve their repositories (Stvilia *et al.,* 2007). For a variety of reasons tools for quality analysis appear not to be widely deployed or used in the IR community. However, as various harvesting projects have shown, there are no significant technical reasons why OAI analysis tools should be unfeasible. In the next sections we outline the design and deployment of two such metadata quality tools.

3. Metadata analysis with KRIS

This section describes KRIS and the *nzresearch.org.nz* website, a metadata aggregation and discovery service. It focuses on the features that help repository administrators measure and improve the quality of metadata.

3.1. Background

KRIS grew out of a collaborative project between The National Library of New Zealand and a group of New Zealand universities and polytechnics. Its goal was to build a national discovery service for the research held in institutional repositories in New Zealand for the mutual benefit of researchers, research users, and research institutions.

The project quickly attracted collaborators from all New Zealand's universities and many of its polytechnics, and launched a New Zealand Institutional Repository (NZIR) mailing list for community discussion. Among their many contributions to the project, these institutions assisted with website requirements and metadata guidelines. The discussion of website requirements included tools to benefit repository managers, and some of these were tools for metadata quality analysis that were subsequently implemented in KRIS.

The metadata guidelines are an integral part of KRIS (National Library of New Zealand, 2007). They are used to promote best practice, consistency and the use of standards in research repositories, and to ensure the discovery service has high quality, nationally consistent metadata. However, the guidelines are also practical and realistic: they prioritise metadata based on how well it supports end-user access, they promote complex metadata but recognise that most repository software will only export unqualified Dublin Core, and they are voluntary (institutions are not penalised for non-compliance). For example, the guidelines are based on Dublin Core, and recommend preferred schemas for Type and Subject metadata,

but also list alternative schemas that will be crosswalked into the preferred schema—and this works (in 99% of cases) even if the metadata is exported as unqualified Dublin Core.

3.2 Measuring Metadata Quality

KRIS has an innovative OAI-PMH harvest framework based on storing three sets of metadata for each record. First, the harvested Dublin Core metadata is stored unchanged. Second, NZIR Internal metadata is generated for each record and used to enhance access to the record by facilitating consistent search and browse across all the participating repositories. It is generated from the harvested metadata using XSL Transformations, and provides metadata for each record in known metadata schemas and controlled vocabularies.

The third set of metadata—and the most interesting for the purposes of this paper—is called NZIR Administrative metadata. This is metadata that describes the quality of the harvested metadata for the record (informally, it is often called "meta-metadata"). Table 1 lists some examples of NZIR Administrative metadata. Each record has zero or more NZIR Administrative metadata fields, and each identifies a specific metadata error, warning, or informational message. An error is defined as a condition that explicitly fails to meet a requirement of the metadata guidelines, such as a required field that is missing. A warning is an example of poor practice that does not explicitly fail a requirement, and informational records are included as advice to administrators (these are discussed in more detail below).

Message type	**Message**
Error	Record has no Author (Creator).
Error	Record has no date
Error	Record has badly formatted date
Error	Record has no Title
Error	Record has no HTTP URL (Identifier)
Warning	Record has no Abstract (Description)
Warning	Author not in "Citename, Firstnames" format
Warning	Type not recognised: Report for External Body
Warning	Type not recognised: Dissertation
Info	Local Type: NonPeerReviewed
Info	Local Type: PeerReviewed
Info	Local Type: Seminar, speech or other presentation

Table 1: Examples of NZIR Administrative metadata including errors, warnings and informational messages.

3.3. Tools

Because the NZIR Administrative metadata quality information is stored in the metadata database like any other metadata, it can be accessed and manipulated as easily as other metadata, and used to build a variety of useful tools.

The primary purpose of the NZIR Administrative metadata is to inform repository administrators about metadata quality issues. One obvious way to do this is to periodically generate a report on the metadata quality and email it to each repository administrator. However, in our planning workshops, the administrators said they did not want that style of feedback—as it results in clogged mailboxes and unread emails.

Instead, the metadata is available "on demand". Metadata errors and warnings are available to administrators when they request it. Several access mechanisms are provided: users can search the collection (or their repository) for metadata errors, or can request the full error set via OAI-PMH export (a specialised *nzir_admin* metadata schema is defined). However, the most popular tool is the RSS feed: any KRIS user can subscribe to an RSS feed of the errors and warnings for a repository (or for the full collection).

Figure 1 shows an example of an RSS feed of errors from the Open Polytechnic of New Zealand institutional research repository, displayed in the Firefox web browser. When this screenshot was taken, there were two records with metadata errors. The browser gives the user the option of subscribing to the feed in their subscription server reader of choice, where they will be notified of new errors as they occur. Clicking the link in each record will take the user to the offending metadata record at the source repository. This mechanism is particularly useful at institutions with self-submission workflows: metadata librarians can monitor the feed for notifications of errors introduced by less experience submitters.

Figure 1. An RSS Feed for metadata errors from the Open Polytechnic of New Zealand research repository displayed in the Firefox web browser.

Another use of NZIR Administrative metadata is to calculate statistics about the metadata quality for each institution, and for the combined collection of records in KRIS—what we call the "state of the nation's metadata." We can use these statistics to compare the performance of different institutions, and can track changes in metadata quality over time. Reports are calculated daily, and users can access the reports at any time. Figure 2 shows a recent KRIS metadata quality report. The final line shows the overall performance: the 5,413 records in the repository contain 35 errors and 337 warnings, which are distributed among 342 different "bad" records. The "state of the nation's metadata" at this point was 93.68% compliant.

Even the relatively simple summary information of Figure 2 shows the important of context in effectively interpreting and using the results from analysis tools. The report could have just shown the percentage of good records for institution and its comparison with the national average. However it is not necessarily the case that an IR with a compliance of 100% is 'better' than one with a compliance of 90%. For example, the former may have only a few tens of records while the latter has thousands. Local understanding of the nature of the institutions, their relative research output and the current progress of their IR in involving departments and researchers will also have an impact on appropriately interpreting such snapshot information. Over time, we may all want to see both the number of records and the percentage compliant to increase, but one-off efforts to increase the former may temporarily degrade the latter.

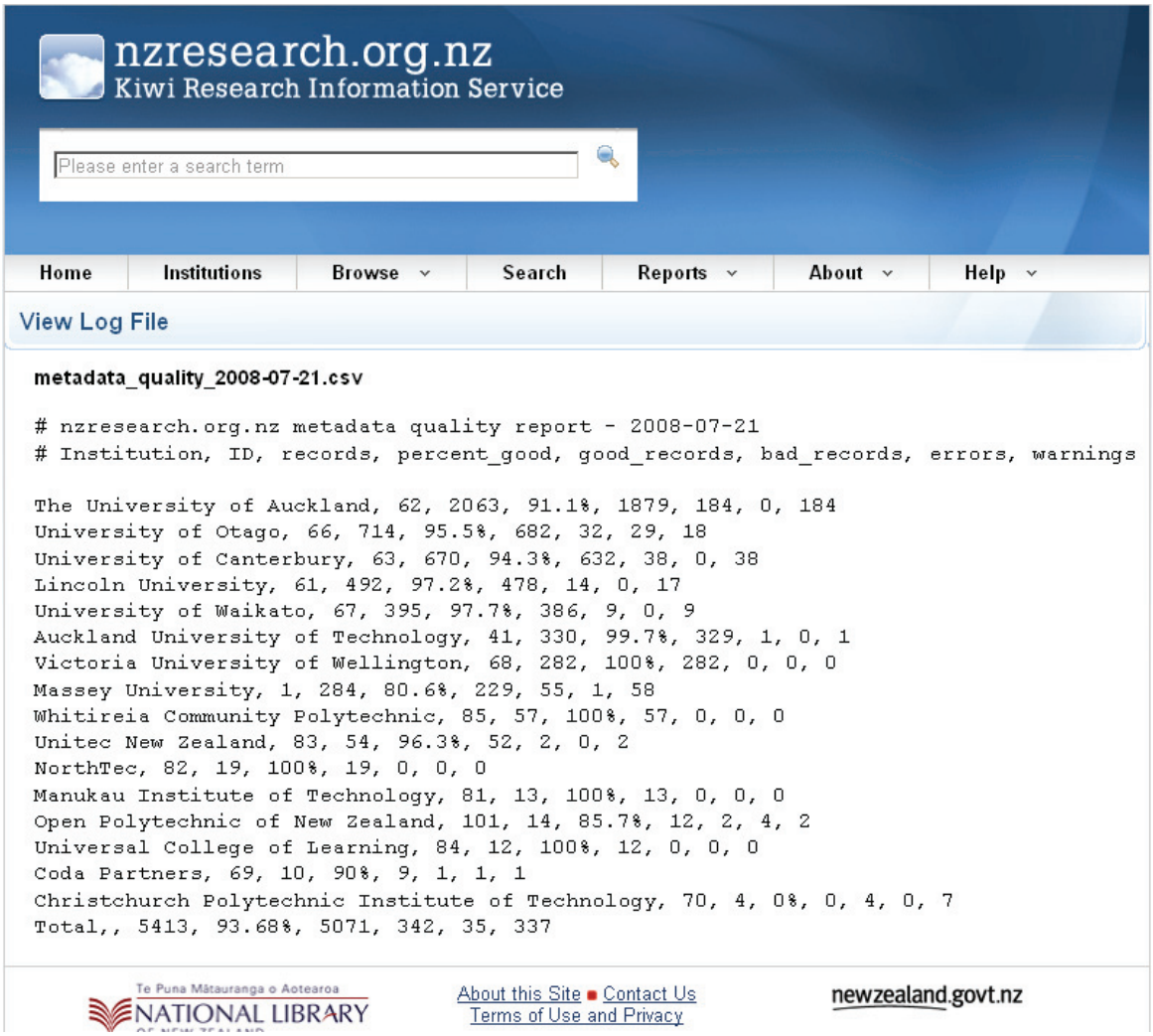

Figure 2. The KRIS metadata quality report for 21 July 2008.

4. MAT tool

This section outlines a web-based metadata analysis tool, MAT, developed alongside the Greenstone digital library tool suite (Bainbridge *et al.,* 2004); for a more detailed description of the tool see Nichols *et al.* (2008).

4.1 Background

The original goal of the MAT tool was to provide a quality analysis component that could be integrated with the Greenstone Librarian Interface (GLI) (Bainbridge, Thompson, and Witten, 2003). Although Salo (2008 to appear) provides valuable insight into the practicalities of running an IR, we found little work that aids software developers understand the needs of repository managers. We chose to build and deploy a prototype tool as the most effective mechanism to solicit user feedback, following the advice of Greenberg and Severiens (2006): "[metadata] tool development needs to be an iterative process between developers and users."

Although GLI is a Java application we chose a Web deployment to reduce technological barriers to use (Golub and Shneiderman, 2003) 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 aimed to allow repository managers to use their own data and so avoid some of the problems of earlier evaluation approaches: "usability of information visualisation 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" (Plaisant, 2004). Thus, our aim was that the prototype would support rapid, incremental requirements capture based on authentic contextualized use.

Technically, the tool is constructed, in a lightweight manner as a servlet in Apache Tomcat embedded in the Greenstone 3 environment. The servlet communicates with existing Greenstone tools for building digital collections and then outputs static HTML quality evaluation reports. Our deployment approach is similar to the *OAI Repository Explorer* service (Suleman, 2001).

4.2 Features of the Analysis Tool

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. Initially, a user enters the URL of an OAI-PMH compliant repository and is then presented with a choice of available metadata prefixes. Once a prefix is chosen the system harvests all the metadata, builds a Greenstone collection, calculates statistics and then presents the user with an HTML report. For IRs with thousands of records this process can take 10 or 20 minutes depending on connectivity and server responsiveness.

The metadata analysis report is structured around the harvested metadata with sections reflecting metadata elements. Figure 3 shows the report for a *dc:title* element with various descriptive statistics and links to further details. This view also shows a sampling of frequency and ASCII sorting, full versions are presented on separate pages. ASCII and frequency ordering were heuristic choices and we expect different types of sorting to be developed as the tool evolves. We have found that unusual or illegal characters often appear at the start or the end of an ASCII sort, as in Figure 5.