http://dx.doi.org/10.1080/01639370902737281

Experiences in Deploying Metadata Analysis Tools for Institutional Repositories

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

SUMMARY. 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, evaluation

---

David M. Nichols and David Bainbridge are Senior Lecturers in the Department of Computer Science at the University of Waikato, Hamilton, New Zealand. Chu-Hsiang Chan developed the MAT tool as part of his MSc studies at Waikato. Gordon W. Paynter is a Technical Analyst at the National Library of New Zealand. Dana McKay is a User Experience Consultant at the Swinburne University of Technology Library in Melbourne, Australia. Michael B. Twidale is Associate Professor at the Graduate School of Library and Information Science at the University of Illinois, Champaign. Ann Blandford is Professor at the UCL Interaction Centre at University College London. Correspondence to David Nichols (dmn@cs.waikato.ac.nz), Department of Computer Science, University of Waikato, Hamilton 3240, New Zealand.

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 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[1] and EPrints.[2] 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.[3] If we accept that "supporting the development of quality metadata is one of the most important roles for LIS professionals,"[4] 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.[5] 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 surveys quality issues for both data and metadata in digital collections, re-iterating that poor quality metadata impedes access to resources.[6] 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 unfeasible. Bruce and Hillmann are explicit: “automated techniques potentially enable humans to use their time to make more sophisticated assessments [of metadata quality]”.[7]

Bruce and Hillmann list seven metadata quality criteria: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility.[8] Criteria such as “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,[9] 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.”[10] 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 but it appears that these surveys have used custom-written software.[11,12,13] 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*.[14]

In summary, there is significant potential for metadata quality tools to allow collection managers to improve their repositories.[15] 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[16] are an integral part of KRIS. 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.

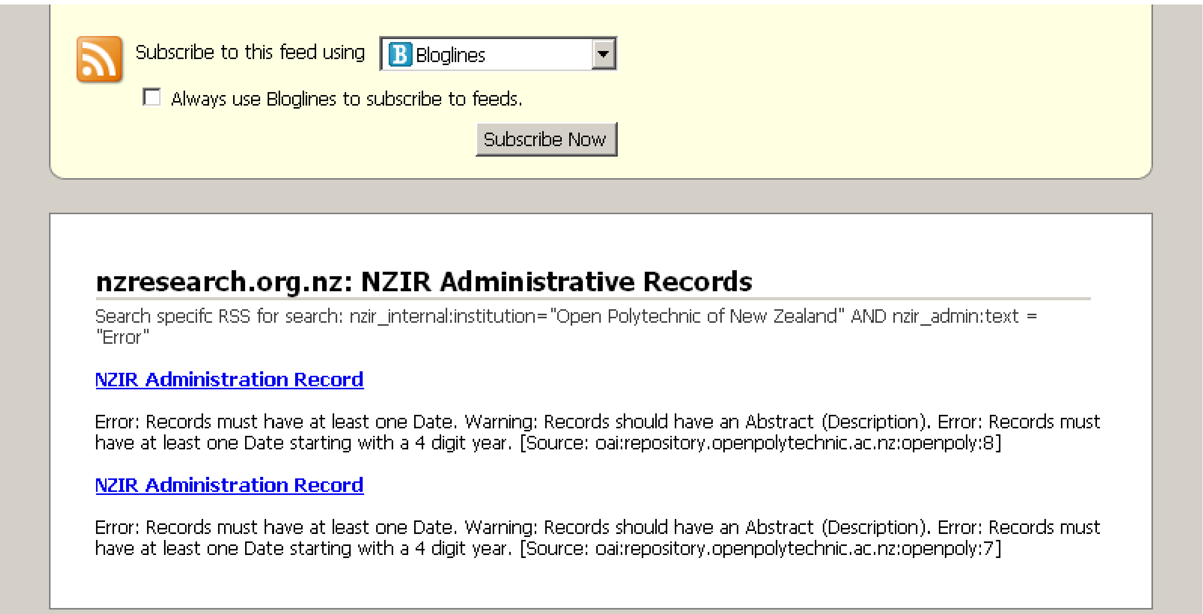
3.3. Tools

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.


Subscribe to this feed using
Bloglines
□ Always use Bloglines to subscribe to feeds.
Subscribe Now

nzresearch.org.nz: NZIR Administrative Records

Search specifc RSS for search: nzir_internal:institution="Open Polytechnic of New Zealand" AND nzir_admin:text = "Error"

NZIR Administration Record

Error: Records must have at least one Date. Warning: Records should have an Abstract (Description). Error: Records must have at least one Date starting with a 4 digit year. [Source: oai:repository.openpolytechnic.ac.nz:openpoly:8]

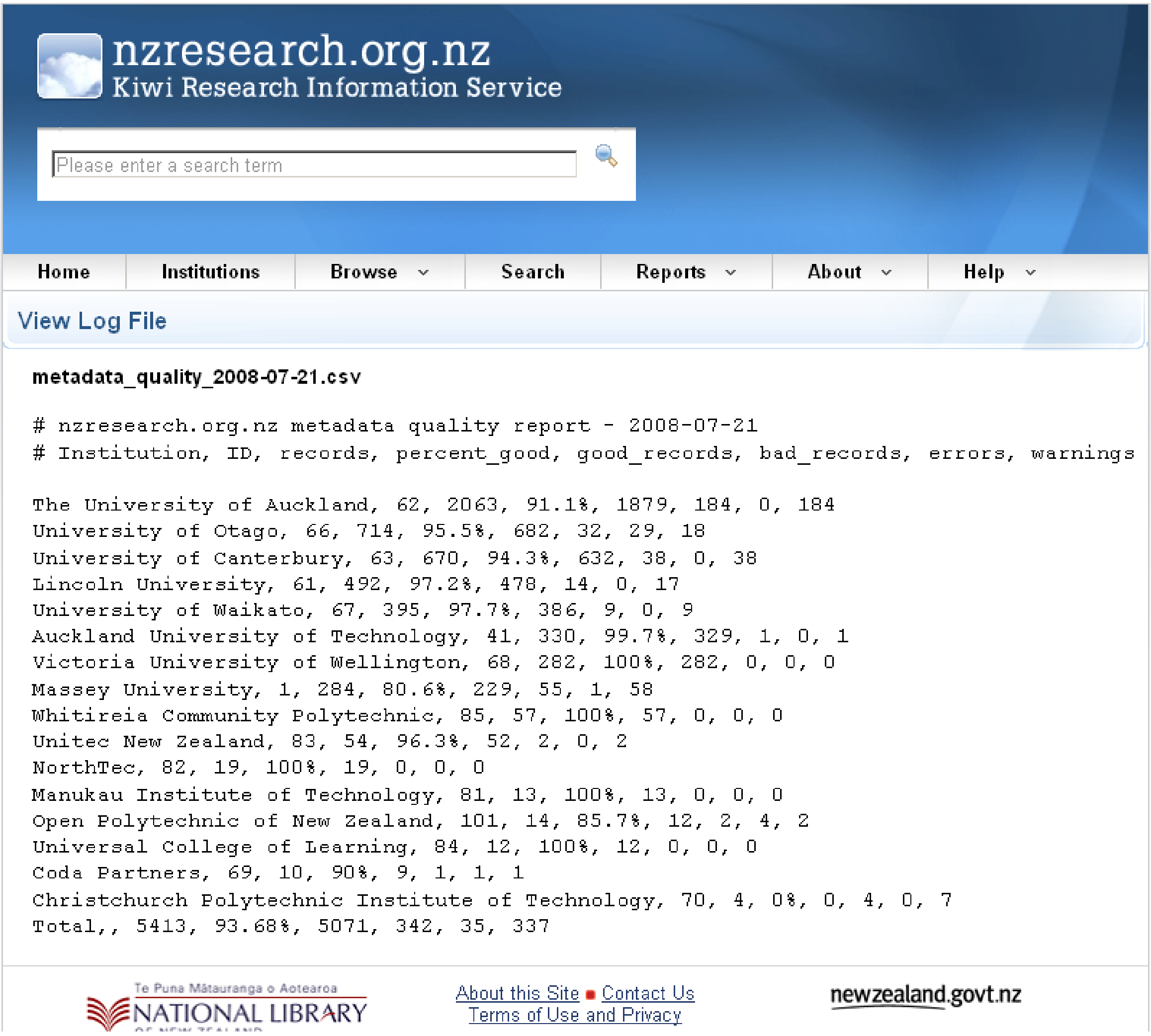
NZIR Administration Record

Error: Records must have at least one Date. Warning: Records should have an Abstract (Description). Error: Records must have at least one Date starting with a 4 digit year. [Source: oai:repository.openpolytechnic.ac.nz:openpoly:7]

Figure 1. An RSS Feed for metadata errors from the Open Polytechnic of New Zealand research repository displayed in the Firefox web browser (July 2008)

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.



metadata_quality_2008-07-21.csv

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

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

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

4. MAT tool

This section outlines a web-based metadata analysis tool, MAT, [17] developed alongside the Greenstone digital library tool suite.[18]

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).[19] Although Salo provides valuable insight into the practicalities of running an IR,[20] 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: "[metadata] tool development needs to be an iterative process between developers and users."[21]

Although GLI is a Java application, we chose a Web deployment to reduce technological barriers to use[22] 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 visualization tools can be measured in a laboratory however, to be convincing, utility needs to be demonstrated in a real setting … Using real datasets with more than a few items, and demonstrating realistic tasks is important."[24] Thus, our aim was that the prototype would support rapid, incremental requirements capture based on authentic contextualized use.

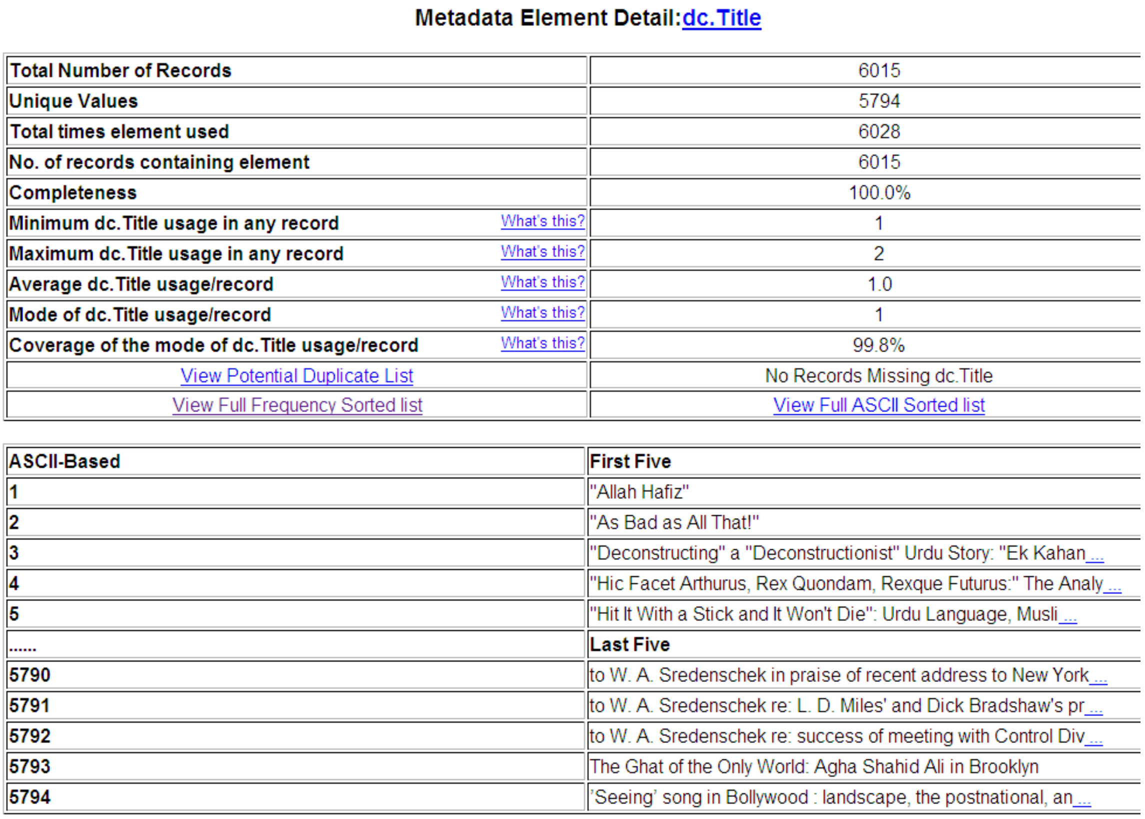
Technically, the tool is constructed in a lightweight manner as a servlet in Apache Tomcat embedded in the Greenstone 3 environment. The servlet communicates with existing Greenstone tools for building digital collections and then outputs static HTML quality evaluation reports. Our deployment approach is similar to the *OAI Repository Explorer* service.[24]

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.

In the case of Figure 3 *dc:title* has 100% completeness, it is defined for every element in the collection, so the link to records without a title is inactive. A list of potential duplicate title values is also provided, using a simple edit distance technique for approximate string matching.[25] Figure 4 shows two sample results for duplicate detection, one an extra space character and one a likely data entry error.[26] We have found that many IR systems appear to lack authority control mechanisms, consequently simple punctuation and spacing differences produce multiple entries reflecting the same person.

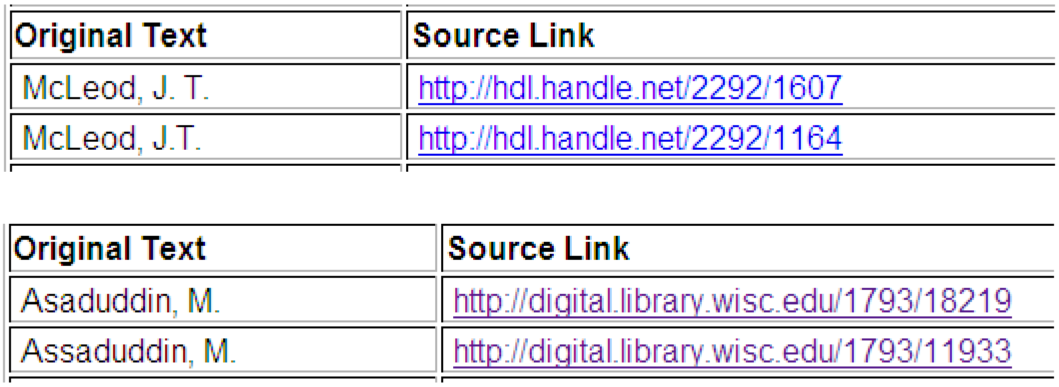


Metadata Element Detail:dc.Title

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

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

Figure 3. The element detail view in MAT



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

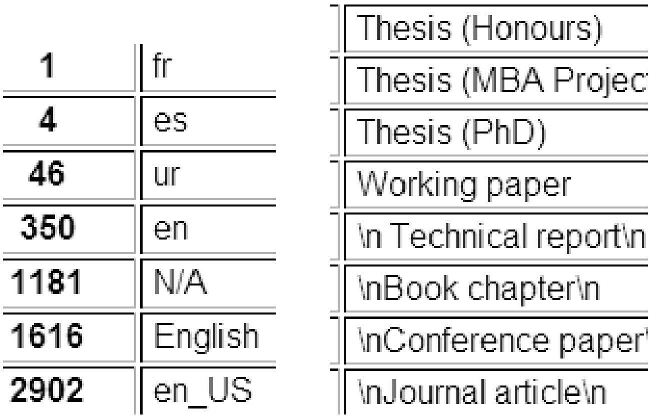
Original Text	Source Link
Asaduddin, M.	http://digital.library.wisc.edu/1793/18219
Assaduddin, M.	http://digital.library.wisc.edu/1793/11933

Figure 4. Two sample results from the duplicate detection report in MAT

The visualisation element of our online tool (Figure 6) closely resembles the example scatter plot of metadata from the Spotfire application described in Dushay and Hillman.[27] 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. MAT has a few simple options to whet users' appetites and try to encourage useful feedback: 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.

Figure 6 shows 13 Dublin Core elements (as two empty elements have been hidden) from 6000 records in a scrolling table. The presence of a metadata item is indicated by a solid 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 the item in the remote repository. The records in Figure 6 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.[28] Although Figure 6 suggests that some values may be placed in the wrong element it requires local knowledge of the repository contents and metadata policies – something which can currently only be supplied by the repository manager. As with much of MAT's output, the issues it identifies are only *potential* problems.

The 6000 records and 13 elements in Figure 6 require an HTML page of about two megabytes, which suggests that visualising significantly larger repositories in this manner may prove to be unfeasible. The links to the source material on the left of Figure 6 require messages to be sent back to a server. This extra communication reduces the size of the web page but has the disadvantage of introducing a dependency; the visualisation links are not available offline.



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

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

Figure 5. Excerpts from an element frequency sort (left) and an ASCII sort (right)

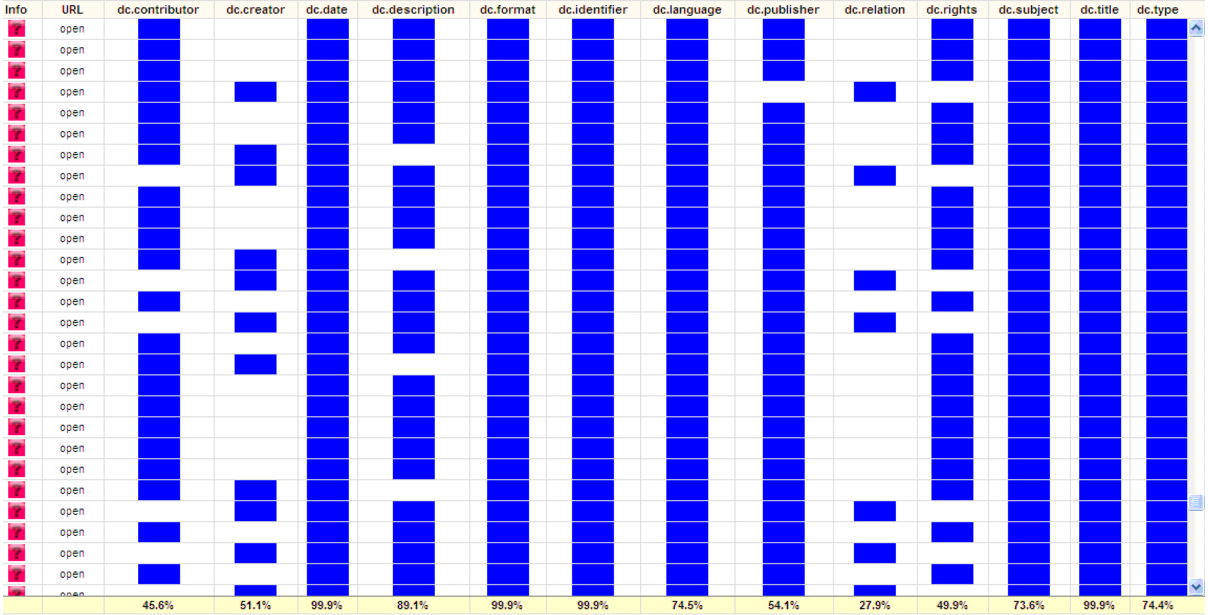


Figure 6. Part of a visualisation of 6000 OAI Dublin Core records from MINDS @ Wisconsin (two empty elements are hidden)

5. Experience and Feedback

We gathered feedback about the tools both informally, through email and conversation, and formally, using online surveys and semi-structured interviews with repository managers. Generally, the remote surveys have been only partially successful in eliciting feedback for improvement, with face-to-face traditional usability think-aloud methods being more useful.

Most feedback received was from repository managers, though some were still planning or developing their repositories. Respondents worked variously with repositories based on Digital Commons,[29] DSpace, Eprints and Fedora,[30] and used many different deposit workflows. Most feedback was generally positive, and has been arranged into logical groups in this section.

5.1 Usefulness and uses

Generally, participants were excited by the tools, seeing a lot of potential for collection improvement (particularly as their own repositories do not offer similar tools). Some repository administrators have consciously used KRIS and MAT to significantly reduce the number of errors in their data.

Repository managers were all asked about the potential uses of MAT, and all said it would be valuable to use MAT to check metadata completeness at periodic intervals. In those cases where feedback comes from someone who has known about the tool for some time, it is apparent that this repeated use actually happens — one survey respondent wrote, “MAT is down, when will it be back up?” Other uses mentioned included checking that an OAI feed was working correctly after a software upgrade, improving metadata entry practices and generating demonstrable metadata quality improvements, and generating statistics not available from their repository software. Interestingly, respondents who were not actively managing a repository found it more difficult to imagine uses for MAT than did active repository managers. However, we also found some interviews about MAT were interrupted, as managers would use their web-based repository administration tools to correct errors they had just found.

One repository manager noted that “Completeness is not an aim: what matters is usefulness.” This is a useful reminder that the information visualisations are highly dependent on informed interpretation, and should not be slavishly followed as a simplistic performance target. Completeness is relatively easy to measure and can be useful in spotting certain problematic patterns in a dataset.[31] But even if a tool identifies some completeness errors, a repository manager may choose to leave them if the cost of correcting them is not justified by the anticipated improvement in usefulness. Equally, a complete dataset may still have distinct usefulness problems. The point is that the tool reports such as visualisations are low cost starting points for informed cost-benefit trade-offs rather than complete solutions to the problem of data quality.

KRIS has more clearly defined uses, and of the tools available, the RSS feeds are by far the most used (though some, like the OAI-PMH feed of errors, are not currently used in any practical way).

5.2 Serendipity

As well as the aggressive metadata cleansing mentioned above, one repository manager mentioned that the 'Top 5' presentations in MAT had allowed her to discover and correct errors while using the tool for another purpose. This serendipitous discovery of errors is likely to be very useful in the less structured metadata environment that MAT faces from its acceptance of any OAI URL as input.

5.3 Result availability

All the KRIS metadata quality reports are publicly available, and can be compared and contrasted (see Figure 2 for example). This openness has encouraged repository administrators to be conscious of the quality of their metadata. Given that anyone can enter any repository URL into MAT and receive a metadata report, this may also be true for MAT, though it was not a use that was mentioned in any feedback. The on-request reports generated by MAT provide an archiving problem as the system has undergone continual evolution in response to user feedback. Consequently, URLs for older reports have become invalid and this has predictably caused problems for some users. The facility for a self-contained static report that could be used independently of the MAT website could address some of these archiving issues.

5.4 Interaction styles

A key feature of both tools is that metadata analysis is available on request, rather than through periodic reports. As a result, the information has to be prepared in advance against possible requests.

However, the two tools have quite different interaction styles overall: KRIS works with a fixed list of repositories, whereas MAT will create reports for any OAI-PMH compliant repository on request. Similarly, KRIS compares the metadata to metadata standards agreed upon by the consortium of represented institutions, generating quite fine-grained feedback, whereas MAT makes few assumptions about metadata standards and reports completeness of all possible unqualified Dublin Core fields.

5.5 Metadata issues

All the repository managers interviewed about MAT were very concerned with metadata quality, and saw value in the at-a-glance depictions of metadata completeness. As one manager commented, “metadata completeness is a mark of record quality.” They were also impressed with the ability to see what kinds of metadata were in their repositories using the list views for individual metadata elements.

MAT was also viewed as an excellent way of checking the quality of metadata translations. All repository managers had been involved in metadata translation from another schema at some point and lacked familiarity with, and tools for, the result (one commented, “I really never think in DC, it’s only used when you need interoperable metadata”).

Currently, all New Zealand research repositories export metadata using unqualified Dublin Core, which places relatively few restrictions on the metadata content. If the metadata were exported in a more complex format, such as a qualified Dublin Core schema using known schemas and controlled vocabularies, then we believe the full potential of KRIS could be realised. For example, the KRIS harvester could check whether metadata values really do comply with their claimed formats and schemas. Having said that, it is important to remember that the primary purpose of KRIS is to support access, and the Dublin Core metadata we are harvesting, though unqualified, is of high quality and supports access well.

On the other side of this, we know of at least one instance where a repository manager was planning to use MAT with qualified Dublin Core metadata, "and that will affect the results." This repository manager reiterated the oft-heard feature request that "it would be great if MAT worked with other metadata formats". Specific requests have been made for METS and MODS; currently these are only partially supported.

Some repositories had metadata in legacy formats that do not match the KRIS metadata guidelines. For example, one has introduced the value *Seminar, speech or other presentation* in the *dc:type* field. In the initial implementation, the use of such unrecognised terms generated a metadata warning, but since they were used consistently this meant that almost every record in that repository was marked "bad". As a result, it was difficult to sort the records with serious metadata problems from those that used a legacy format, and the KRIS tools were therefore not useful to the repository administrator. We therefore added a new category of administrative metadata, the "informational" message, that is not considered a "bad" record, but which does note and discourage non-compliance. Some examples are shown in Table 1.

5.6 Technical and deployment issues

The OAI-PMH provides good support for metadata analysis. In KRIS, metadata quality information is created and updated for each record as it is harvested, so is always current (as at the most recent harvest) and available for reporting and access. By using nightly incremental OAI-PMH harvests, KRIS can maintain a full set of metadata for quality review without performing full harvests. Any metadata errors that are fixed at the source repository in response to feedback from KRIS (or MAT) will be re-harvested overnight, and the NZIR Administrative metadata will be re-generated to incorporate the changes. While the same efficiencies are theoretically possible in MAT, the use of Greenstone as an underlying tool means that results are not available until the metadata has been harvested, the statistics recalculated, and incurring a much longer delay. In the current implementation, MAT reports are typically regenerated on demand and scheduled harvesting is not yet available.

Several deployment issues with KRIS were ironed out in the first few months of use. For example, we have observed that when KRIS finds an error with the URL in the *dc:identifier* field, it can be difficult to refer the user back to the source record. We have fixed this by noting the OAI-PMH identifier in the description field of the RSS output (see Figure 1). Configuration can also be quite labour-intensive: when we want to highlight new metadata issues or errors, we have to update the XSL file that is used to generate NZIR Administrative metadata from the harvested Dublin Core. Both KRIS and MAT implementers concluded that better monitoring of real-world usage would sometimes have helped with understanding and highlighting when the tools behaved unexpectedly.

5.7 Feature requests

All respondents had some feature requests. Some wanted MAT to deal with different types of metadata, some asked for more documentation (others missed features during an initial exploration), and many wanted links from the sorted element views to the associated documents so they could immediately repair incorrect metadata. Usability improvements and documentation are clearly a priority for further development. Most would like the tool to work faster to build the reports. In at least two cases feedback was from users who noticed and appreciated changes that had been made in line with these feature requests, saying, for example, that “the interface is still simple but I noticed you can click right through to the original record now, and that is *so* useful!”

KRIS users made similar feature requests to MAT users, for example, in early versions all the KRIS tools for repository administrators were located in different parts of the website; this made them hard to find and use. We have now introduced a “repository details” page that links to all the useful tools.

6. Metadata analysis tool design

Based on our experience with KRIS and MAT, we offer the following advice to developers of future metadata quality review tools:

- What problems are you trying to solve? Understanding the metadata? Looking for specific problems? You should have a specific repository administrator problem in mind before you start.
- Who are your users? Repository administrators are end-users, frequently metadata specialists, seldom technologists. Integrated help / tutorial content will be necessary.
- Do not assume that repository administrators know their OAI URLs and/or have control over what is harvested from their repositories. These are technical issues, not metadata issues.
- Well thought out visualisations are considered useful by repository managers, and much appreciated. However, some managers expressed strong preferences for the textual and statistical approaches, which suggest that both forms of presentation should be available.
- Web-based tools can work especially efficiently when a web-based IR administrative interface is available as well. Your analysis tool should provide links from every item of interest back to the record in the source repository to facilitate error correction.

- Agreed metadata policies help with buy-in, as does working in a well-defined community of repositories with a common goal.
- Exploit the advantages of OAI-PMH. Overnight, zero-effort updates to the analysis are appreciated.
- Will your results be private (to repository administrators) or publicly accessible? Metadata in institutional repositories seldom has restrictions on its use to the extent that the described content does, and if the metadata itself is available to all-comers via OAI-PMH, then there is no reason to restrict access to the analyses.
- There are many existing OAI-PMH compliant tools that you can leverage, but do not be surprised if, when adapting or reusing existing software, (Greenstone, in the case of MAT) you stress it in unusual ways and uncover constraints that normal use may not encounter.
- There are potential security implications when deploying a tool like MAT that allows users to nominate a site to harvest. You may be giving external users the ability to bypass firewalls and security restrictions. Site like KRIS that harvest a fixed list of source repositories are safer.

7. Conclusion

MAT and KRIS are examples of metadata analysis tools for institutional repositories that harvest metadata using the common OAI-PMH protocol, analyse the harvested metadata, and provide tools and visualisations that help repository administrators to understand their metadata, and to improve it. They are both available to repository administrators in New Zealand, who have used them to increase the metadata quality of their research outputs. Formal and informal feedback shows these tools are useful and well received. Neither tool was particularly complex or expensive to develop, and yet the relatively simple analyses that they provide were found to be helpful in a range of settings and uses. The feedback obtained from actual use can be used in the development of more sophisticated functionalities and improvements to the interface.

We hope these tools, and others like them, will continue to be used to improve metadata and also the processes for its creation and subsequent refinement. We note that there are trade-offs in the establishment, growth and development of IRs. Initially, learning by individuals and rapid growth to attain a critical mass may require compromises in metadata quality. This is not necessarily a problem; a meticulously accurate but very limited collection may not be as useful as a large or near-complete collection containing many errors. Metadata quality analysis and visualisation tools can help repository administrators make informed decisions about these trade-offs and how to best allocate resources to manage overall quality and size.

Even providing such tools as MAT and KRIS introduces another trade-off: repository administrators with limited resources must decide how much of their limited time to spend analyzing and correcting problems in existing metadata records, and how much to dedicate to describing new material to add to their repository. We expect that different repositories will make different trade-offs at different times, and we hope that tools such as these can help inform the process.

References

1. Robert Tansley, MacKenzie Smith and Julie Harford Walker. “The DSpace open source digital asset management system: challenges and opportunities,” in *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005),* LNCS 3652. (Berlin: Springer, 2005), 242-253.

2. EPrints (2008) Software. http://www.eprints.org/software/

3. Dorothea Salo, “Innkeeper at the Roach Motel,” *Library Trends* 57 no. 2 (2008): 98-123. doi: 10.1353/lib.0.0031

4. R. John Robertson, “Metadata quality: implications for library and information science professionals,” *Library Review* 54 no. 5 (2005): 295-300. doi: 10.1108/00242530510600543

5. Sarah L. Shreeves, Ellen M. Knutson, Besiki Stvilia, Carole L. Palmer, Michael B. Twidale, and Timothy W. Cole, “Is quality metadata 'shareable' metadata? The implications of local metadata practices for federated collections,” in *Proceedings of the Twelfth National Conference of the Association of College and Research Libraries*. (Chicago, IL: Association of College and Research Libraries, 2005), 223-237.

6. Jeffrey Beall, “Metadata and data quality problems in the digital library.” *Journal of Digital Information* 6 no. 3 (2005), Article No. 355, 2005-06-12, http://journals.tdl.org/jodi/article/view/jodi-171

7. Thomas R. Bruce and Dianne I. Hillmann, “The continuum of metadata quality: defining, expressing, exploiting,” in *Metadata in Practice*, Dianne I. Hillmann and Elaine L. Westbrook, eds. (Chicago, IL: American Library Association, 2004), 238-256.

8. Ibid., p. 243.

9. Besiki Stvilia, Les Gasser, Michael B. Twidale and Linda C. Smith, “A framework for information quality assessment,” *Journal of the American Society for Information Science and Technology*, 58 no. 12 (2007): 1720-1733. doi: 10.1002/asi.v58:12

10. Naomi Dushay and Dianne I. Hillmann, “Analyzing metadata for effective use and re-use,” in *Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2003)*, (Seattle, WA, 2003). http://hdl.handle.net/1813/7896

11. Miles Efron, “Metadata use in OAI-Compliant Institutional Repositories,” *Journal of Digital Information* 8 no. 2 (2007) http://journals.tdl.org/jodi/article/view/196/169

12. Shreeves *et al.*, Is quality metadata 'shareable' metadata? The implications of local metadata practices for federated collections.

13. Jewel Ward, “Unqualified Dublin Core usage in OAI-PMH data providers,” *OCLC Systems & Services* 20 no. 1 (2004): 40-47. doi: 10.1108/10650750410527322

14. Hussein Suleman, “Enforcing interoperability with the open archives initiative repository explorer,” in *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '01),* (New York, NY: ACM, 2001), 63-64.

15. Stvilia *et al.*, A framework for information quality assessment.

16. *National Research Discovery Service Metadata Guidelines*. National Research Discovery Service Project, National Library of New Zealand, 2007. http://www.natlib.govt.nz/catalogues/library-documents/national-research-discovery-service-metadata-guidelines

17. David M. Nichols, Chu-Hsiang Chan, David Bainbridge, Dana McKay and Michael B. Twidale, “A lightweight metadata quality tool,” in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL’08),* (New York, NY: ACM, 2008), 385-388.

18. David Bainbridge, Katherine J. Don, George R. Buchanan, Ian H. Witten, Steve Jones, Matt Jones and Malcolm I. Barr, “Dynamic digital library construction and configuration,” in *Proceedings of the Eighth European Conference on Research and Advanced Technology for Digital Libraries (ECDL'04),* LNCS 3232, (Berlin: Springer, 2004), 1-13.

19. David Bainbridge, John Thompson and Ian H. Witten, “Assembling and enriching digital library collections,” in *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2003),* (Washington, DC: IEEE Computer Society, 2003), 323-334.

20. Salo, Innkeeper at the Roach Motel.

21. Jane Greenberg and Thomas Severiens, “Metadata Tools for Digital Resource Repositories: JCDL 2006 Workshop Report*,” D-Lib Magazine* 12 no. 7/8 (2006). http://www.dlib.org/dlib/july06/greenberg/07greenberg.html

22. Evan Golub and Ben Shneiderman, “Dynamic query visualizations on World Wide Web clients: a DHTML solution for maps and scattergrams,” *International Journal of Web Engineering and Technolology,* 1 no. 1 (2003): 63-78.

23. Catherine Plaisant, “The challenge of information visualization evaluation,” in *Proceedings of the Working Conference on Advanced Visual interfaces (AVI '04).* (New York, NY: ACM, 2004), 109-116.

24. Suleman, Enforcing interoperability with the open archives initiative repository explorer.

25. Gonzalo Navarro, “A guided tour to approximate string matching,” *ACM Computing Surveys* 33 no. 1 (2001): 31-88. doi: 10.1145/375360.375365

26. Beall, Metadata and data quality problems in the digital library.

27. Dushay and Hillmann, Analyzing metadata for effective use and re-use.

28. Ibid.

29. Digital Commons (2008) Software. http://www.bepress.com/ir/

30. Carl Lagoze, Sandy Payette, Edwin Shin, and Chris Wilper, “Fedora: an architecture for complex objects and their relationships,” *International Journal on Digital Libraries* 6 no. 2 (2006): 124-138. doi: 10.1007/s00799-005-0130-3

31. Bruce and Hillmann, The continuum of metadata quality: defining, expressing, exploiting.