Quality Assessment of Metadata in Open Archives

Bhavana Harrilal
University of Cape Town
Cape Town, South Africa

Abstract

High quality metadata allows for effective and valuable digital libraries. A system which harvests and processes metadata was built to assess metadata quality. The system showed that metadata quality can be determined automatically. Implementing such an assessment allows for metadata to be measured and standardised effectively, overall improving the quality and efficiency of digital libraries.

Introduction

The quality of metadata is a key factor in the efficiency and dependability of digital libraries. Metadata that is of poor quality delays resource discovery and the organization of electronic resources. The need for metadata of a good quality ensures interoperability and digital identification. These two components allow systems to exchange and identify data without loss of content and functionality. Metadata quality has a direct impact on services provided to users, such as the ability to search for items based on metadata. Digital information is fragile therefore high quality metadata ensures the longevity and accessibility of resources. The assessment of metadata quality thus allows for a digital library's efficiency.

Background

The research papers on this topic before took a different approach to measure quality. These are a statistical approach, a conceptual framework, identifying quality characteristics or detecting quality problems. The paper by Bui & Park [1] implemented the Dublin Core schema as a basis of assessment. They used the NDSL repository to build a system that converted metadata into a spreadsheet via Excel, employing chosen fields that were weighted to give a higher assessment of quality. This statistical approach showed that the quality of metadata depended on the collection that was being tested. The paper entitled "A conceptual framework for metadata assessment" [4] showed a new conceptual framework for metadata quality and a method for its assessment that exploits logic rules which are interdependent with the metadata. Another statistical by Baden [2] implemented an algorithm that scored each metadata record on a scale (0-10) based on adherence

to DC and OLAC domain-specific controlled vocabularies. Using this value and derived values, an assessment of quality can be made. A combination of human evaluation (qualitative) and automatic evaluation (quantitative) was adopted by Drexel University to conduct the quality assessment of metadata of the Internet Public Library (IPL) [3]. This qualitative method gave an indication of the quality of information by rating accuracy, completeness, consistency and functionality. The quantitative method only measured the completeness of metadata in the collection.

Design of solution

Considering the effect metadata has on a digital library, a system that harvests and processes metadata to gauge its quality was implemented. By harvesting metadata the system runs a quantitative quality analysis of the inputted archive, which divides an archive then divides that into records where metadata for each record can be assessed. This was done in two ways: the validity of the metadata elements within a record and the elements correspondence to the recommended standards i.e. ETD-MS (Interoperability Metadata Standard for Electronic Theses and Dissertations) and DC (Dublin Core). Using these standards as a measure of quality, a scale of quality can be implemented:

- If the element exists then quality points are awarded. These elements are:

title	subject	creator
description	publisher	contributor
date	type	format
Identifier	language	coverage
rights	thesis.degree – this has sub-elements (i.e. name, level, discipline, grantor)	

- If the element value is relevant, additional points are given. This is determined by format restrictions or length restrictions, given by ETD standards. Mandatory elements as defined by ETD-MS were given larger weights if the element adheres to the ETD-MS standards.
- The maximum number of points per element is 5.
- Using the points system a value can be calculated to give a record an individual metadata quality assessment.
- Using a scale of assessment as follows to define quality:

Quality Value	Quality Assessment
X<40	Bad
40<X<45	Weak
45<X<55	Average
55<X<60	Good
X>60	Very Good

- Using this system that evaluates a record as a whole the entire archive may be analysed in a similar fashion. Such data will give an indication so as to the quality of the archive, its validity and correspondence to the recommended standards namely, ETD and DC. Here the strongest and weakest elements in the archive can be singled out. An average of the archive quality may also be assessed and rated on the scale of assessment.

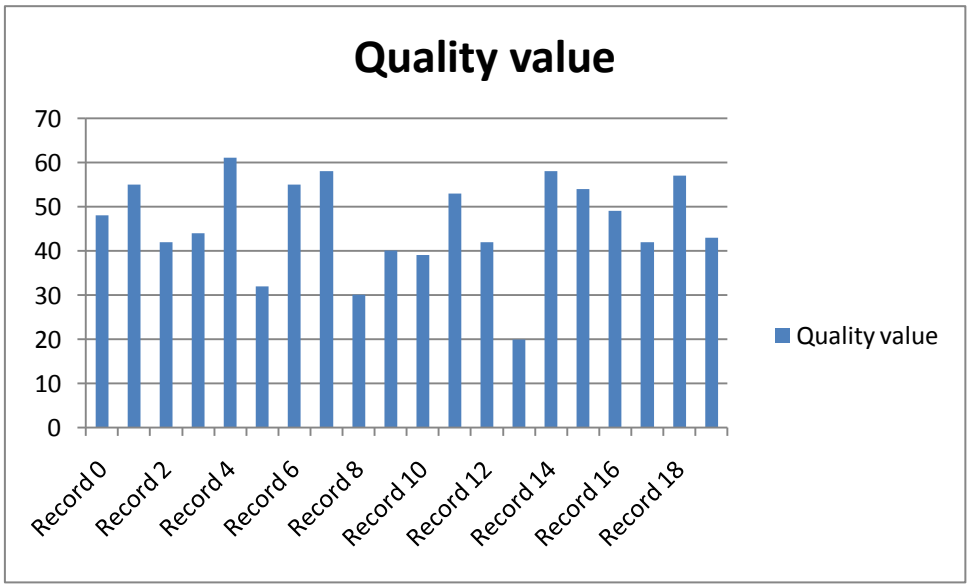
Experiments

An archive of data was harvested using the Harvey Perl harvester that implements an OAI harvester. This captures records into an XML document, which the system processes and evaluates.
This XML document can then be. Here tests can be run which indicate strengths and weakness of the metadata, either as individual records or the collection of records.

As a test to assess whether the application quantitatively correctly assesses the metadata correctly a sample of 20 metadata records were processed. These were selected to test if the application was effectively and accurately rating each metadata record in accordance with the quality scales defined earlier. The sample was chosen so as to assess quality of all ranges of potential metadata quality, that being metadata of a very good quality to those which had a very bad quality.

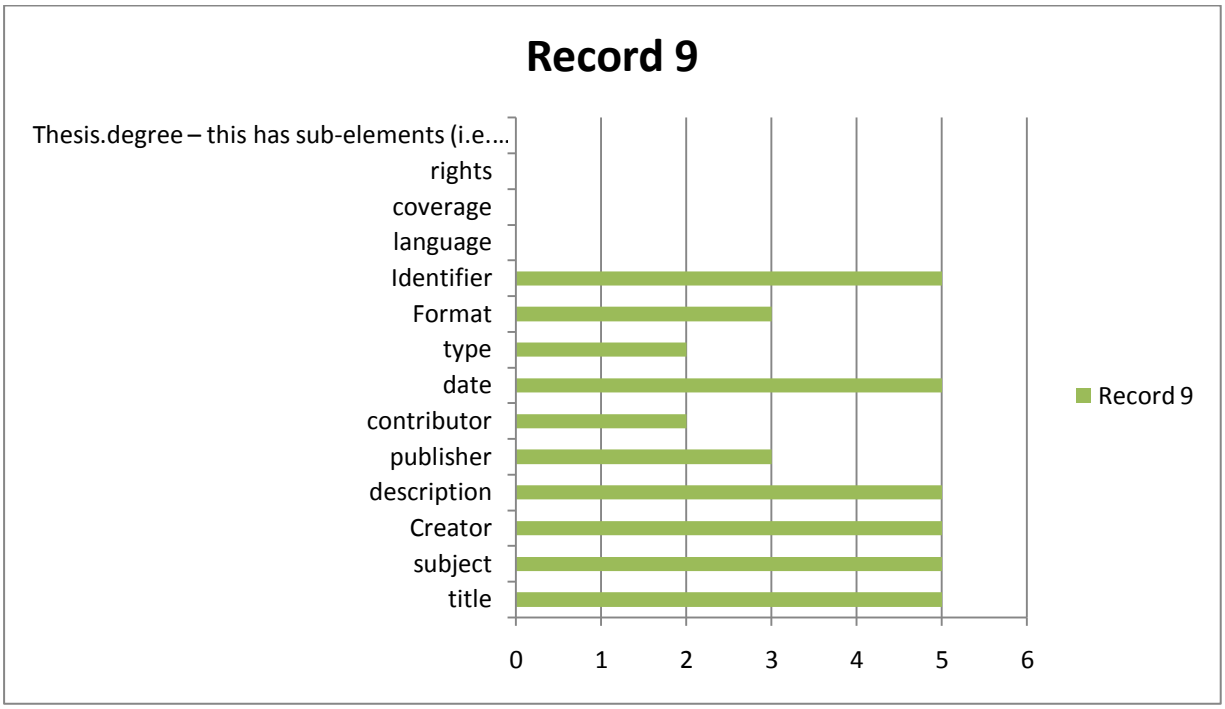
Results

The results concerning the sample space of 20 records showed that indeed the application does test according to the defined quality assessment scale accurately with the following results:



Each record could then be further analysed to indicate the weak points and strengths of the metadata record

Take Record 9 for example:



This graph shows the weak points and the lack of information present judged in accordance with the ETD-MS standards.

Using the functions to analyse each metadata record the same can be evaluated about a collection of metadata.

A collection in the form of an XML document was processed and the following results were found:

```
Analysis of Metadata Quality Assessment:

Collection Assessment

Collection total: 14145
Average quality value: 47
AVERAGE overall metadata quality.

Element Assessment:

dc:subject        Strong element
dc:description    Strong element
dc:publisher       Average to strong element
dc:contributor    Weak element
dc:date  Strong element
dc:type    Average to strong element
dc:format         Weak element
dc:identifier     Strong element
dc:language       Weak element
dc:coverage       Weak element
dc:rights         Weak element
Weakest element is dc:identifier
Strongest element is dc:coverage
```

Using output such as this, quantitative conclusions can be made based on the assessment that has taken place.

Conclusion

This paper shows that by using the ETD-MS and DC standards a definition of quality can be established and used to assess metadata successfully. The system was built to solely assess metadata; this system can calculate both a single record and a collection of records which can be assessed. The quality is measured quantitatively, by the existence and likeness to the standard for thesis metadata established by NDLTD. By the ability to assign a quality value to the record this shows quality of metadata can be assessed automatically.

Future work

There is room for improvement as a more generic solution can be established. Currently the system supports Vanilla XML encoding. The integration of MARC-21 encoding could help to provide a more general solution. Reworking the quality value scales can assign values differently, taking into account different standards. Also it is possible to focus quality on a different specification e.g. specified metadata element/s. The system looks purely at the metadata alone and does not assess the metadata XML schema - it may be possible to assess quality using this schema.

References

1. Bui, Yen, and Jung-ran Park (2006). An assessment of metadata quality: A case study of the national science digital library metadata repository. In Haidar Moukdad (ed.), CAIS/ACSI 2006 *Information Science Revisited Approaches to Innovation* from http://www.caisacsi.ca/proceedings/2006/bui_2006.pdf.
2. Hughes, Baden. (2004). Metadata quality evaluation: Experience from the open language archives community. *Digital Libraries: International Collaboration and Cross-Fertilization* from http://www.springerlink.com/content/4kaxeu5p2fb2nac1
3. Ma, Shanshan, Caimei Lu , Xia Lin and Mike Galloway. (2008) . Evaluating the metadata quality of the IPL from http://www.asis.org/Conferences/AM09/openproceedings/papers/49.xml
4. Margaritopoulos, Thomas, Merkourios Margaritopoulos, Ioannis Mavridis, and Athanasios Manitsaris. (2008). A Conceptual Framework for Metadata Quality Assessment. From http://dcpapers.dublincore.org/ojs/pubs/article/download/923/919

5. DCMI Usage Board. (2008). DCMI Metadata Terms. Retrieved January 28, 2011, from http://dublincore.org/documents/2008/01/14/dcmi-terms/
6. Dublin Core Metadata Element Set, Version 1.1. (2008). DCMI Metadata Terms. Retrieved January 28, 2011, from http://dublincore.org/documents/2010/10/11/dces/
7. Guidelines for implementing Dublin Core in XML. (2003). Powell, Andy, and Pete Johnston. Retrieved February 2,2011, from http://dublincore.org/documents/2003/04/02/dc-xml-guidelines/
8. ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations version 1.00, revision 2. (2008). Atkins, Anthony, Edward Fox, Robert France, and Hussein Suleman. Retrieved January 19,2011, from http://www.ndltd.org/standards/metadata/etd-ms-v1.00-rev2.html