Metadata and Data Quality Problems in the Digital Library

by Jeffrey Beall

jeffrey.beall@cudenver.edu
University of Colorado at Denver and Health Sciences Center,
Downtown Denver Campus

Abstract

This paper describes the main types of data quality errors that occur in digital libraries, both in full-text objects and in metadata. Studying these errors is important because they can block access to online documents and because digital libraries should eliminate errors where possible. Some types of common errors include typographical errors, scanning and data conversion errors, and find and replace errors. Errors in metadata can also hinder access in digital libraries. The paper also discusses the responsibility for errors in digital documents and offers suggestions for managing digital library data quality.

1. Introduction

Data quality is important in the digital library because high quality data insures accurate and complete access to online objects and because users expect and deserve accurate, error-free data. Most studies of erroneous data in the context of libraries have focused on online library catalogs. But as libraries make more and more content available digitally, the issue of data quality in digital objects will increase in importance. This article focuses on both metadata errors and errors in the actual documents and summarizes the issues and possible solutions related to typographical errors and other types of data quality problems in the digital library.

In the context of digital libraries, there are three levels or perspectives of data quality. The first level is *absolute data quality*, which refers to the overall level of data quality of both digital objects and metadata within a digital library. The second level of data quality is *faithful reproduction data quality*, and this refers to the data quality of objects that originated elsewhere, that is, outside the digital library. Faithful reproduction means that objects in a repository reproduce exactly the documents or objects as they were in their original form. The third level of data quality is *born digital data quality* and it refers to the quality of data in the digital library for objects and metadata that were born digital within the individual digital library. This level measures the data quality of everything produced by an individual digital library.

2. Related Literature

The two aspects of digital library data quality are the quality of the data in the objects themselves, and the quality of the metadata associated with the objects. These two, digital data quality and digital metadata quality have been studied in many diverse contexts and communities for varying functions besides information retrieval, such as digital preservation (Rothenberg, 1996), data used in simulations and models (Rothenberg & Rand, 1997), databases (Medawar, 1995), and in museums (Marty and Twidale, 2000). Considerably more has also been written about the quality of metadata, a discussion that builds on studies of data quality in online library catalogs and in descriptive cataloging (Graham,

1990; Massey, 2003) for retrieval than about digital data quality. There is also literature about the quality dimensions of both information (Wormell, 1990) and data (Fox, et al, 1994). This paper reviews only selected studies that are directly pertinent to digital data and metadata quality in an operational digital library, which is broadly construed to include full-text abstracting and indexing databases as well as the associated metadata.

Ojala (1996) describes retractions, corrections, and amplifications in the online environment. These activities become necessary when errors are found in online documents, chiefly research papers. Also, when published research or reporting is later found to be fraudulent, publishers and digital libraries have to determine what action to take regarding the articles, i.e., whether they should be removed from the servers or retained, or whether they should be retained, but with a notice indicating they were later found to be false.

Lesk (2005) points out the need for having authoritative texts in digital libraries. He writes,

> "For many humanities researchers, it is important to have accurate, well-edited texts online; versions adequate for popular reading are not sufficient. Insufficient evaluation has been made of the editorial quality of some online texts, particularly those contributed by volunteers. And in some cases the desire to avoid copyright problems has resulted in the use

> of nineteenth-century texts instead of modern and more accurate texts still under copyright protection" (p. 58).

In other words, what Lesk is describing is the need for faithful reproduction data quality in digital libraries.

An early study of spelling errors in scholarly, machine-readable documents was completed by Pollock and Zamora (1983). They found a rate of misspelling of 0.2% and found that for words containing a spelling error, 90-95% of them have only a single error. Although this study was undertaken before the advent of sophisticated spell-check software, a comprehensive data quality typology of errors distinguishing among omissions, insertions, substitutions, transposition, and multiple errors is used to characterize and help in the analysis of the misspellings investigated.

3. A Taxonomy of Data Quality Errors

3.1 Typographical Errors

One of the many advantages that digital objects have over their print counterparts is that once an error is fixed in the digital document, it is fixed forever. Errors in printed works, of course, last as long as the physical item. We've all seen (and perhaps made) typographical errors in online documents, such as web pages. In the online environment, the problem of typographical errors is probably underestimated because information searchers don't realize when a typo has prevented access to a document. They likely assume that the

search results represent a complete and accurate retrieval based on their search criteria, when in fact, dirty data may be causing some relevant objects to be excluded.

A brief article by Gardner (1992) provides a simple classification of spelling and typographical errors. She lists them as:

1. Errors of letter omission
2. Errors of letter insertion
3. Errors of letter substitution
4. Errors of letter transposition

A more scientific approach to typographical errors that draws on the principles of psycholinguistics is provided by Berg (2002). He studied "500 typographical errors excerpted from scholarly works published in English" (p. 187). The errors he studied were the hardest to discover and correct, for they had persisted throughout the high level of editing and scrutiny that generally occurs in scholarly journals. He found that most errors involved single letters.

Gentner et al. (1983) present a "Terminology for errors." Table 1 is a summary of the error categories they list, along with definitions and examples.

Table 1

Error category	Definition	Example
Misstrokes	errors traced to inaccurate motion of the finger	[None given]
Transposition errors	two consecutive letters in a word are interchanged	typing *iknd* for *kind*
Interchange errors	two non-consecutive letters are interchanged	typing *jamor* for *major*
Migration errors	One letter moves "migrates" to a new position	typing *atht* for *that*
Omissions	a letter in a word is left out	typing *omt* for *omit*
Insertions	an extra letter is inserted into a text	typing *asnd* for *and*

Substitutions	occurs when the wrong letter is typed in place of the correct letter	[None given]
Doubling errors	A word containing a repeated letter is typed so that the wrong letter is doubled	typing *bokk* for *book*
Alternation errors	A letter alternates with another, but the wrong alternation sequence is produced	typing *thses* for *these*

Table 1. Typographical error categories, definitions of the categories, and examples of typographical errors. Table adapted from Gentner, et al. (1983)

One argument minimizing the problem of typographical errors suggests that because a misspelled word is likely to appear again—correctly--in the same document, little chance exists that the error will hinder access. While this may be true in some cases, typographical errors can still hinder access for phrase and proximity searching, and typos can create greater obstacles when they occur in the metadata associated with an object, for that metadata often serves as a surrogate for the entire object. Indeed, for many digital objects such as images and documents whose full text is not indexed, the metadata is the only access point the search interface searches to retrieve them. In this case, the data quality of the metadata is a crucial element of accurate retrieval.

Figure 1 shows an example of a typographical error found in an online document. In this case, there is a typographical error in the title of the article. The word “and” is misspelled.

Incentives for Increasing Return Rates: Magnitude Levels, Response Bias, amd Format

J. SCOTT MIZES, E. LOUIS FLEECE, AND CINDY ROOS

THE use of monetary and nonmonetary incentives has received significant attention as a method of increasing return rates from mail surveys. Armstrong (1975), in a review of eight studies, concluded that monetary incentives have a substantial impact on survey return rate, typically reducing nonreponse by a third. Nonmonetary incentives also have positive effects, although the response increase has not always been statistically significant (cf. Nederhof, 1983). Initial evidence from direct comparisons suggests that monetary incentives are superior to nonmonetary ones in increasing mail survey returns (Goodstadt, et al., 1977).

Armstrong's review concluded that increasing the size of a monetary incentive increases mail survey response. However, he notes that few studies have compared different magnitudes of monetary incentives, and few studies have investigated the effects of incentives greater than a dollar, leaving the effects of larger incentives unknown. Thus, the current study was designed to address these issues by comparing the effects of $5 and $1 incentives on mail return rate.

Abstract The use of monetary incentives has been shown to significantly increase response rate. However, previous investigations have rarely investigated the effects of incentives greater than $1, compared different magnitudes of incentives, or investigated response bias due to incentives. The current study also investigated the utility of an Answer Check. Results suggest that monetary incentives increase response rate, larger incentives do not necessarily further increase survey response, incentives do not appear to bias responses, and the Answer Check does not facilitate response rate.

J. Scott Mizes is Assistant Professor, Department of Psychology, North Dakota State University, Fargo, ND 58105. E. Louis Fleece is Assistant Professor of Psychiatry (Psychology), University of Alabama—Birmingham School of Medicine, and Psychologist, Birmingham VA Medical Center. Cindy Roos is a graduate student, Department of Psychology, University of Alabama—Birmingham. This research was supported by a grant to J. Scott Mizes from the Department of Psychiatry, University of Alabama—Birmingham School of Medicine.

Public Opinion Quarterly Vol. 48:794–800 © by the Trustees of Columbia University
Published by Elsevier Science Publishing Co., Inc. 0033-362X/84/0048–794/$2.50

Figure 1

Figure 1 (Mizes, Fleece & Roos, 1984, p. 794) is an example of a typographical error in an online document. In this example, the word “and” is misspelled in the title.

Typographical errors can also occur at the word or sentence level. For example, a word can be left out of a sentence, changing its meaning. This omission is especially serious when the omitted word is the word “not.” Additionally, sentences or paragraphs can be left out, and people can be misidentified. Also, an image may be labeled incorrectly, be mixed up with another image, or have erroneous metadata associated with it.

In terms of digital libraries, probably the most serious typographical error is the one that occurs in a URL. These errors can completely block access to a document. But with effective digital library management, URL errors can be caught and fixed.

3.2 Scanning and Data Conversion Errors

A relatively new type of error occurs in digital objects: scanning errors. This type of error occurs when scanning hardware or software incorrectly renders the text from printed to digital format. Scanning software sometimes puts spaces in the middle of words, and it can incorrectly read an individual letter in a word, such as misrepresenting the letter “l” for the letter “i.” Many such errors can be eliminated by a careful editing process that employs human review or a spell check function, but automated processes are not foolproof. It is possible to observe scanning errors by searching typographical errors in a full text database such as J-STOR.

If the typographical error cannot be found in the documents retrieved, it is likely that the error did not occur in the original document but in the scanning process. A common scanning error is the mis-rendering of the word “that” as “tbat.” To test this, one can search the word “tbat” in a database of scanned text, such as J-STOR. In most cases, the search will not turn up text containing “tbat” because the word was never in the text in the first place; it was generated from a scanning error.

It’s also possible for errors to creep into text documents when they are converted from one format to another. For example, it’s possible for errors to occur in the conversion of a document from Microsoft Word format to HTML. Without proper editing, these character conversion errors sometimes remain in the archived object in the digital library and are problematic because they affect the indexing and retrieval of the document and because they are not a faithful reproduction of the original. In other words, they have low faithful reproduction data quality. This type of error tends to occur more often with letters from languages other than English, languages with diacritics, and with symbols.

3.3 Find and Replace Errors

While a find and replace error generally occurs in the preparation of original documents, it can occur with digital library objects during a conversion process. One recent example of this type of error occurred in *The Journal of Academic Librarianship* in 2003 (Schottlander, 2003). The author of a book review

mentioned a librarian named *Charles Husbands*. But during the production phase of the journal issue a find-and-replace algorithm changed the name to *Charles Spouses*, a change which appeared in the final print and online published versions of the journal issue. This is an example of poor born digital data quality.

4. Metadata Errors

Metadata errors in digital libraries can occur in many forms. Where metadata errors exist, they can easily block access to material available through a digital library. These errors are most serious when metadata serves as a surrogate for objects held in a digital library and full text searching is not available. Image databases are particularly vulnerable to metadata errors because virtually all search access to image databases is through metadata. The importance of metadata cannot be overstated. According to Guy, Powell, and Day (2004) “there is an increasing realisation that the metadata creation process is key to the establishment of a successful archive.” Similarly, Robertson (2005, p. 295) states “Supporting the development of quality metadata is perhaps one of the most important roles for LIS professionals.”

Moen, Stewart, and McClure (1998) list the different aspects of metadata quality:

> Access, Accuracy, Availability, Compactness, Comprehensiveness, Content, Consistency, Cost, Data structure, Ease of creation, Ease of use, Economy, Flexibility, Fitness for use, Informativeness, Quantity, Reliability, Standard, Timeliness, Transfer, Usability.

Basically, this list tells us where problems can occur with metadata. Metadata that is deficient in any of these areas can and does affect resource discovery. The authors concede, however, “Schemes inevitably represent a state of compromise among considerations of cost, efficiency, flexibility, completeness, and usability … ” (Moen, Stewart, and McClure, 1998, p. 248).

Bruce and Hillmann (2004) attempt a similar analysis of metadata quality. Their measures include completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. But they warn that “Most metadata communities outside of libraries are not yet at a point where they have begun to define, much less measure, quality” (p. 240).

Some metadata is created automatically, by means of special software programs. This process can lead to errors in the metadata and to failed searches because metadata creation generally needs human intervention to be successful. For example, even the most sophisticated computer program might not be able to differentiate among locks (hydraulic engineering) or locks (hardware) or locks of hair, air locks, etc., or among authors with similar names. Artificial intelligence has not progressed to the point that it can be substituted for the human analysis, interpretation, and classification of digital objects.

Crosswalking metadata from one scheme or format to another can also be a source of errors. While sophisticated programs exist to crosswalk metadata from

Dublin Core to MARC, for example, the process is not foolproof. Crosswalking metadata errors are more serious when the data is converted from a less rigid metadata scheme (such as Dublin Core) to a scheme where data values are more tightly controlled (such as MARC).

Metadata harvesting, often promoted as a cost-effective tool for aggregating metadata and providing access to a broader range of digital information resources, is also problematic. Errors can occur in the actual harvesting, such as data transmission errors, and errors can crop up with incompatible data elements or formats. Eclectic metadata can be corrupted when it is converted to a common scheme. After harvesting metadata from multiple sources a digital library may be faced with metadata of varying structures, content standards, quality, and schemes that make it inconsistent, unreliable and all but unusable.

5. Responsibility for Error

With the exception of metadata, an agency involved in the reproduction of documents is not responsible for the data quality errors they contain. In other words, we do not expect digital libraries to spend resources correcting errors in its documents that originated elsewhere. Indeed, it is often the responsibility of the digital library to preserve the documents in their original state. One exception to this is when a producer of an original document reissues that document, such as when a journal publisher or an author issues an erratum for a previously published issue, or when the publisher or author simply corrects an error in a

document. Digital libraries need to have a mechanism or a policy that allows these types of corrections to trickle up to their archived digital objects. Alternatively, digital libraries have the option of storing the original document and the replacement, but there should be a clear relationship established in the metadata for both documents.

6. Example and Analysis of Typographical Errors

To exemplify the problems of erroneous data in a full-text online database, I searched five common typos from the website "Typographical Errors in Library Databases" (Ballard, 2005) in JSTOR. JSTOR is a repository of searchable, scanned images of journals. The selection of the search terms was not random; I selected five words containing typos that I felt would show the significance of the problem of errors in a digital repository. Table 2 shows the results of my searches.

Word containing typographical error	Number of hits
artic	355
enviroment	163
managment	265
offical	451
univeristy	961

Table 2. The number of hits in JSTOR for five common typographical errors. [Searched in March, 2005].

It's not only important to point out the extent of these errors; their etiology is also worth noting. First, they could be errors that originally appeared in the print version of the journal. JSTOR includes many journals that began publication

decades ago, before the advent of computers and spell-check software. Second, they could be errors that originated in the digitization process, namely the scanning errors mentioned earlier. The errors that occurred in the original print version cannot be fixed in the online version because the online version is ideally a faithful reproduction of the original pages of the journal. The pages with these errors show high faithful reproduction error quality but low absolute data quality. The solution of the problem of bad data in original documents must lie elsewhere, namely in the retrieval software.

Measuring data quality in digital libraries is made difficult by the fact that documents are constantly being added to them. The size is not static. Moreover, it is difficult to compare data quality among separate databases because they are of various and often unknown sizes, and because their sizes change almost daily. So, although it is difficult to generate a rate of error for any given dynamic, online database, one might use sampling to estimate the error rate. For example, it might be possible to measure the number of errors per megabyte of data.

7. Managing Data Quality and Typos in Digital Libraries

Developing strategies for dealing with typographical errors and other data quality problems in digital libraries can lead to an improvement in data quality and user access to digital objects. The best strategy is to prevent errors from occurring in the first place. Improved editing of original documents will lead to a reduction of error rates in the documents. Libraries and other consumers of digital data need

to begin to demand data with a lower rate of error from database producers. The high cost of commercial online databases justifies the expectation of error-free data. Also, libraries and database vendors need to expand their efforts to develop search software that simultaneously searches known misspellings of words along with their correct spellings. Such a system would allow for more complete search retrievals and allow for a greater access to archived digital documents.

There are four ways for managers of digital libraries to handle data quality errors in digital objects found in digital libraries they are:

- Fix the error in the document
- Make available a new document that replaces the document with the error
- Make available a new document that contains a notice of the error and its correction (this document would be hyperlinked to and from the original document containing the error)
- Use special search software to compensate for some errors

Fixing an error in a document is appropriate when the digital library is responsible for the content of the document. That is, the digital library authored the document or bears intellectual responsibility for it. For metadata, it is always appropriate for a digital library to correct metadata errors. This is true whether the metadata was created by the digital library or whether the metadata was harvested or acquired from an external source.

Making a new document available that replaces a former one is similar to publishing a revised or corrected edition of a print work. The digital library usually has the option of keeping or eliminating the earlier version of the document. If the decision is to keep the older, outdated document, hyperlinks should be added that link both versions of the documents. The distinction between the two should be clearly identified in the metadata for each document, and possibly also within the documents themselves.

A third way to deal with data quality errors is to issue or append short and separate documents that describe and correct the error or errors in a separate earlier document. This practice follows the example of some scientific journals in which the author or authors of an article write a short article to correct errors that appeared in an earlier article. For digital libraries, the disadvantage of this practice is increased resources devoted to arranging the earlier document and its subsequent corrections. But the advantage would be the ability to provide both the original document (the one with the errors) and the corrected document. In the context of digital libraries, sometimes corrections are appended to the end of the original document and don't exist as a separate document but instead exist as an appendix to the original one.

Finally, it may be possible for the search software to compensate for the typographical errors that exist within a digital library. The Google search interface has a mechanism for dealing with typographical errors at the point of searching.

When a user inputs a search that contains a typo, the interface supplies hypertext links that search the correct form of the misspelled word. A search interface that does the opposite of this might be a workable solution to the problem of typos in digital libraries. In other words, whenever a user searches a correctly spelled term, the search interface would provide the option of pulling up documents that contain misspelled versions of that word. Alternatively, this searching of misspelled words could be done in a way that is invisible to the user.

8. Conclusion

Clean data always bears a high cost. But in the context of digital libraries, the benefit of this cost is accurate, error-free data and consistent access to that data. Data quality control is an essential part of digital library management. As the amount of digital information continues to increase, the management of data quality in digital libraries will not only continue to be one of the more important aspects of digital library administration, but additional research and investigation also becomes critical.

Further research into two areas of digital library data quality would likely prove valuable. First, the development of a standardized method for calculating and comparing data quality among different databases would help digital library managers measure data quality and focus on data that needs remediation. Second, more research into the error rate of scanning of textual objects is needed. Research is needed to determine whether the error rates of optical

character recognition are acceptable and to what extent they hinder searching and document access.

References

Ballard, T. (2005) *Typographical Errors in Library Databases*. Rev. Jan. 20, 2005. Online: http://faculty.quinnipiac.edu/libraries/tballard/typoscomplete.html

Berg, T. (2002) "Slips of the Typewriter Key", *Applied Psycholinguistics*, Vol. 23, 185-207.

Bruce, T.R. and Hillmann, D.I. (2004) "The Continuum of Metadata Quality: Defining, Expressing, Exploiting". In *Metadata in Practice*, edited by D.I. Hillmann and E.L. Westbrooks (Chicago: American Library Association) pp. 238-256.

Gardner, S. (1992) "Spelling Errors in Online Databases: What the Technical Communicator Should Know", *Technical Communication*, Vol. 39, 50-53.

Gentner, D.R., Grudin, J.T., Larochelle, S., Norman, D.A., Rumelhart, D.E. (1983) "A Glossary of Terms Including a Classification of Typing Errors". In *Cognitive Aspects of Skilled Typewriting*, edited by W.E. Cooper (New York: Springer-Verlag), pp. 39-43.

Graham, P. (1990) "Quality in Cataloging: Making Distinctions", *Journal of Academic Librarianship*, Vol. 16, 213-218.

Fox, C., Levitin, A. and Redman, T. (1994) "The Notion of Data and its Quality Dimensions", *Information Processing and Management*, Vol. 30, No. 1, 9-19.

Lesk, M. (2004) *Understanding Digital Libraries*, 2nd ed. (Boston: Elsevier)

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

Marty, P. and Twidale, M. (2000) "Unexpected Help with Your Web-based Collections: Encouraging Data Quality Feedback from your Museum Visitors", *Museums and the Web 2002 Papers.* Online: http://www.archimuse.com/mw2000/papers/marty/marty.html

Massey, O. (2003) *Auditing catalogue quality by random sampling*. A master's dissertation submitted in partial fulfillment of the requirements for the award of Master of Arts degree of Loughborough University. August 2000. Available online: http://owen.massey.net/dissertation/index.html

Medawar, Katia (1995) “Database Quality: A Literature Review of the Past and a Plan for the Future”, *Program,* vol. 29, no. 3, 257-272.

Mizes, J.S., Fleece, E.L., Roos, C. (1984) “Incentives for Increasing Return Rates: Magnitude Levels, Response Bias, amd [sic] Format”, *Public Opinion Quarterly*, Vol. 48, No. 4, 794-800.

Moen, W.E., Stewart, E.L., and McClure, C.R. (1998) “Assessing metadata quality: findings and methodological considerations from an evaluation of the U.S. Government Information Locater Service (GILS)”. In *IEEE International Forum on Research and Technology Advances in Digital Libraries, ADL ’98 : Proceedings, April 22-24, 1998 Santa Barbara, California* (Los Alamitos, Calif.: IEEE Computer Society Press) pp. 246-255.

Ojala, M. (1996) “Oops! Retractions, Corrections, and Amplifications in Online Environments”, *Searcher*, Vol. 4, No. 1, 30-41.

Pollock, J.J., and Zamora A. (1983) “Collection and Characterization of Spelling Errors in Scientific and Scholarly Text”, *Journal of the American Society for Information Science*, Vol. 34, No. 1, 51-58.

Robertson, R.J. (2005) “Metadata Quality: Implications for Library and Information Science Professionals”, *Library Review*, Vol. 54, No. 5, 295-300.

Rothenberg, J. (1996) “Metadata to Support Data Quality and Longevity”. Paper presented at the 1st IEEE Metadata Conference, Silver Spring, MD.

Rothenberg, J., & Rand (1997) “A Discussion of Data Quality for Verification, Validation, and Certification (VV&C) of Data to be Used in Modeling”, Rand Project Memorandum PM-709-DMSO, Rand. See also Data Quality Templates: http://vva.dmso.mil/Templates/Dataquality/default.htm

Schottlander, B.E.C. (2003) “Metadata fundamentals for all librarians” [Book review], *The Journal of Academic Librarianship*, Vol. 29, issue 6, 418-419.

Wormell, I., editor (1990) *Information quality: Definitions and dimensions* (Los Angeles: Taylor Graham)

Glossary

Absolute data quality
The overall level of data quality of both digital objects and metadata within a digital library

Born digital data quality

The quality of data in the digital library for objects and metadata that were born digital within the individual digital library

Crosswalking
The mapping of data elements or content from one metadata scheme to another

Data conversion errors

Errors that occur when data is converted from one format (such as HTML) to another (such as Microsoft Word format)

Database vendors
Businesses or organizations that sell proprietary data to libraries

Dirty data
Data that contain errors

Faithful reproduction data quality
The data quality of objects that originated elsewhere, that is, outside the digital library

Find and replace errors
Errors that are created in a document when a find and replace algorithm does not work as intended

J-STOR
A proprietary database that consists of scanned images of journals

Metadata errors
Errors that occur in metadata

Metadata harvesting
The aggregation of metadata generally from one or more external sources

Phrase searching
Searching for a phrase, such as “Statue of Liberty”

Proximity searching
Searching for a word or phrase that occurs close to another word or phrase in a document

Scanning errors
Errors that occur when text from a physical object such as a book are scanned and converted to a digital object

[end]