An Assessment of Google Books' Metadata

Ryan James, University of Hawai'i Manoa, Honolulu, Hawaii
Andrew Weiss, Delmar T. Oviatt Library, The California State University,
Northridge, California

Abstract

This article reports on a study of error rates found in the metadata records of texts scanned by the Google Books digitization project. A review of the author, title, publisher, and publication year metadata elements for 400 randomly selected Google Books records was undertaken. The results show 36% of sampled books in the digitization project contained metadata errors. This error rate is higher than one would expect to find in a typical library online catalog.

(Published in the *Journal of Library Metadata* 12(1) 2012)

Keywords

Google Books, metadata, electronic books, digital library

Author Contact information

Ryan James, University of Hawai'i at Manoa Libraries, 2550 McCarthy Mall, Honolulu, HI 96822. E-mail: rsjames@hawaii.edu

Introduction

The quality of Google Books' metadata is an area of concern for scholars and librarians. Google Books is the largest book digitization project currently in progress. The immense size and ambition of the project results in an unprecedented amount of attention, but also highlights one of its major problems: its inability to instill trust that Google Books records and Google Books Search results accurately reflect the contents of the whole catalog.

Pope and Holley (2011) show that despite being open access, public domain books are accessible under varying conditions ranging from "full view" to "limited" and "snippet" previews to those without any previews available. Contrary to fears brought up by critics, notably libraries and publishers, Google Books has expressed its desire to work with such institutions in order to function as a "virtual card catalog of all books in all languages that helps users discover new books and publishers discover new readers" (Pope & Holley, 2011). As a result, Google Books' stated goals need to be examined thoroughly within the prism of library best practices.

Researchers investigating the Google Books project are primarily focusing on the quality of manuscript images and optical character recognition (OCR) as well as the prevalence of metadata errors. Townsend (2007) and Duguid (2007) each focus on the problems of OCR and scanning found in the Google Books Project. A previous article by James (2010) on the Google Books project assesses the quality of scanning as it related to a text's legibility; in a review of 2500 pages from 50 randomly selected books, it is determined that about 1% of pages contained errors of legibility. Additionally, Nunberg (2009b) describes the numerous errors found in Google Books' metadata. The list of errors includes: incorrect publishing dates, classification errors, altered titles, misattribution of authors, and links to the wrong texts. Stopping short of providing specific rates-of-error statistics while calling it a "mishmash wrapped in a muddle wrapped in mess", Nunberg believes error to be "endemic" to the Google Books experience (Nunberg, 2009b).

A representative of Google, Jon Orwant, in response to an August 29, 2009, blog post by Nunberg (2009a), explains that the metadata was not derived from an automated computer process, but was gathered from several metadata providers. This process, presumably, involves using humans to generate the metadata. Additionally, he argues that when Google Books has lacked a BISAC category, "we guess correctly about 90% of the time" (Nunberg, 2009a). The resulting situation is explained more as an aggregation of previously created human error than any true culpability on the part of Google Books. Yet the issue remains that Google has "tried to master a domain that turned out to be a lot more complex" than first thought, and the company has not paid sufficient attention to the quality of Google Books' metadata (Nunberg, 2009a).

These issues raise some important questions: For example, is a 10% rate of error in missing BISAC categories acceptable? What is an acceptable rate of error for the other types of errors shown by Nunberg, James and others? To look at these with a better context, it is important to look at previous studies of traditional library services.

Ryans (1978), Ballard and Lifshin (1992), Chapman and Massey (2002), and Beall (2005) provide important analyses of the role that errors take in reducing findability in bibliographic data systems ranging from OPACs to Digital Asset Management Systems. Misspellings have traditionally been the focus of most studies; Ryans (1978) focuses on the most frequent cataloging errors; Ballard and Lifshin (1992) find 1000 spelling errors among 117,000 keywords (a rate of about .85%); Chapman and Massey (2002) find an overall error rate of 34.4% in their study of errors in MARC records at the University of Bath, but most of the errors are considered "minor". Only 7.6% - 11.4 % of errors are considered "major," or "those that affect access" (Chapman & Massey, 2002). These studies seem to suggest that anywhere between 1% and 12% rates of major errors would be comparable to the standards found in libraries' rich, metadata-centric systems.

Many of these studies, however, emphasize “mechanical accuracy” over that of “subjective intellectual accuracy” (Chapman & Massey, 2002). This study takes another approach by focusing not merely on mechanical accuracy such as misstrokes, transpositions, interchange and migration errors, but also looks at some of the problems involved with preserving semantic meanings in metadata: omissions, insertions, misattributions, and other factual errors, including the frequency in the misuse of dates. These can be argued as major errors as well, since they impede the ability of researchers to find the actual records not only in terms of mechanical retrievability, but also semantically.

Methodology

A word was randomly selected from the Oxford English Dictionary and searched in Google Books to generate a results list (Oxford English Dictionary Online, 2010). The list was then limited to preview and full view. Next a random number was generated between 1 and 100 and was used to select a record to examine from the results list (Random.org, 2010). Words generating results lists of fewer than one hundred books were discarded and a new results list was generated from a new word.

The scanned pages of the selected book records were examined to identify the title, author, publisher, and publication date. Books for which this information could be determined were compared to their corresponding item record in Google Books. Books for which this information could not be found or definitively ascertained were discarded. 472 books and their metadata records were examined using this method, 72 of which were discarded because it was not possible to verify or ascertain the all of the metadata elements relevant to this study.

The title, author, publisher, and publication date metadata fields were evaluated as either being free of error, given the constraints of this study, or containing at least one error per record examined. This true or false test evaluation of metadata fields does reduce the ability of this study to describe the nuances and specific types of errors encountered.

This methodology’s chief advantage is that multiple books can be chosen randomly for examination. This helps to control for biases resulting from certain suppliers of metadata providing systematically poor quality metadata to Google Books, and instead offers a fairer assessment of the overall quality of the metadata in Google Books.

This methodology focuses on books that Google Books provides scans of in the form of “full view” and “limited preview”. With these books often scans of the title pages, verso, and colophon were available for inspection. From these scanned pages the accuracy and existence of the metadata elements for author, title, publisher, and publication date could be determined.

It was not possible to gather the same data from books that were in “snippet view” and those that had no scanned pages available for inspection, so they were excluded from this study.

Data were gathered during July and August of 2010. Since one person coded the data, there are no statistics on intercoder reliability.

Results

The author, title, publisher, and publication metadata fields were examined for 400 books, providing a sample with a margin of error of +/- 4.9% at a 95% confidence interval. A total of 203 errors were found, with 31 title errors, 48 author errors, 83 publisher errors, and 41 publication date errors. 36.75% (+/ - 4.9% at a 95% confidence interval) of the books sampled in the study had metadata errors. The average error per book was 1.97. Three books had four errors, seven had three errors, 33 had two errors, and 104 had one error. 147 books had at least one error. Most common were errors relating to the attribution of publishers, followed by author, publication date, and title.

Summary of Results		
Metadata Field	Errors Found	% of Errors
Publisher	83	41%
Author	48	24%
Publication Date	41	20%
Title	31	15%
Total	**203**	**100%**

Publisher Field:

The largest amount of errors was located in the Publisher field. The most common types of errors for this field were related to the complexity of the publishing industry. Many of the errors (26 of 83) in this field occurred because of misattributing the parent company as publisher rather than the subsidiary company, which was the actual publisher. Other errors include listing the wrong publisher and providing no publishers at all.

Author field:

Errors in the author field occur mainly because Google Books conflates the role of author, editor, and translator and treats them as the same intellectual entity. This is problematic from a semantic point of view. While all parties surely play an important role in the production of the book, Google Books lumps them into the category of author, making no differentiation for their respective roles. Using Google's broad definition of authorship, 48 errors were found.

Publication Date Field:

The most common types of errors for the publication date field were mainly due to providing the wrong year. Unfortunately for users, determining the publication date is complicated by the messiness of what an "edition" is versus a "reprinting". Multiple copyright dates also cause problems. Within standard library cataloging practices this problem can be solved by listing multiple dates, yet Google Books often chooses to give only one date.

Title field:
Among the most common of the problems for the title metadata has been the misattribution of the title, providing the wrong title, and truncating a long sub-title or complex main title.

Discussion

Regardless of the source of the metadata errors, Google Books is the end point of a supply chain that delivers metadata to the user. As such, the onus is on Google Books to ensure the metadata they provide is as reasonably accurate as possible. Google Books' explanation for the origin of the errors is informative, but does not alter its responsibility to provide accurate metadata to the user (Nunberg, 2009a).

The overall error rate of 36.75% found in this study suggests that Google Books' metadata has a high rate of error. While "major" and "minor" errors are a subjective distinction based on the somewhat indeterminate concept of "findability", the errors found in the four metadata elements examined in this study should all be considered major.

The author and title metadata errors are probably more troubling to the general user, as these errors likely have a larger impact on a general user's search of the Google Books database. Yet, even if one discounts publisher and publication date errors as minor, it only reduces the instance of major errors to 19.75%. A nearly 20% instance of error is still two to three times the amount of major errors found in the University Of Bath's catalog, which ranged between 7 and 12% and included a more rigorous standard for spelling errors than this study (Chapman & Massey, 2002). It could even be argued that with more attention paid to spelling errors in Google Books among the author and title fields that the rate of error might be much higher than 19.75%.

However, despite the greater impact of author and title on the general user and given Google Books' explicitly stated desire to be the "virtual card catalog" of the online digital world, the publisher and publication date element errors should be considered "major" as well, given their importance in determining copyright issues and the accurate verifying of a text's edition. The 31% error rate for these two elements should be cause for concern for all users, however, and not just those with specialized information needs, as it points toward a general disregard for the importance of metadata. Accurate versions and textual dates are important considerations not only for scholars in the humanities, who often rely on the analysis of multiple versions of a text to discuss writers and their works, but they also provide an anchor to physical objects with which even the general user will need in order to develop a sense of trust with the collection. Without verifiable editions, Google Books will be unable to provide the reliable information needed by scholars and general users alike.

Furthermore, the problem with Google Books' approach to metadata is that it is misleading. Scholars interested in the work of a particular publishing house will find this metadata unhelpful. Even more troubling, the lack of verifiable publisher metadata decreases transparency about who actually holds copyright for a given book. Given the frequent acquisitions and mergers occurring in the publishing industry, one would be right to question if Google Books accurately assigned the publisher. This also seems to go against its original intention of helping "publishers discover new readers" (Pope & Holley, 2011).

If Google Books truly does want to become the virtual card catalog, then it will need to prove that it has not just poorly reinvented the wheel. It will need to find ways to improve its metadata so that it can even compete with existing library systems. Google Books' main advantage, an emphasis on the full-text keyword search, which Google has both pioneered and refined, cannot currently make up for the problems in the metadata. The rates of error for the four metadata fields looked at in this study are much too high for users to feel comfortable that they will be able to find all that they are searching for.

Google's approach to building a digital library appears to favor quantity over quality. The advantage to this approach is they have rapidly scanned millions of volumes. The disadvantage is many of their records have faulty metadata and a small, but significant number of their books contains poorly scanned pages (James, 2010).

One can surmise that, intentionally or not, Google is relying on full text searching of key words to compensate for faulty metadata records. This is a functionality the traditional library card catalog does not have, and does promise some benefit to the user. The key word approach fails, however, in the specific ordering of search results list and in the general concept of findability.

We do not know the inner workings of the proprietary algorithms Google Books uses to order the search results list, but we can see that metadata is featured prominently on the search results list. A user may be misled by faulty metadata displayed on the search results list, such as an incorrect author or title, into not reviewing relevant books. A user may also attempt to sort the search results list by publication date, and due to faulty metadata exclude relevant books.

When we think more broadly on the concepts of recall and precision, and also on the user's perception of recall and precision, the conflict between faulty metadata and keyword searching takes on larger proportions. If a user is judging the effectiveness of their search based upon a search results list constructed through a combination of keyword searching and metadata, and in which the metadata is prominently featured on the search results list, any errors in the metadata would have a disproportionate effect on the user's perception of recall and precision. The likely end result is that faulty metadata will obscure relevant results, whether through the specific ordering of the search results list or by discouraging users to review certain books.

In the debate over the relevance of metadata in the era of full text searching, this article can offer no firm conclusions. However, from the user's perspective, faulty metadata could be a barrier to access.

It remains to be seen how much effort Google will put into correcting their faulty metadata and poorly scanned pages in Google Books. The error rates found in this article suggest they have some ways to go before being comparable to a typical library catalog.

The purpose of this article is not to be anti-Google Books, but to explore and provide data on some of the problems with Google Books' metadata. Google Books is currently the largest digital library, and by examining where it stumbles and comes up short, valuable lessons can be learned that might improve not only Google Books, but also the numerous other digital libraries currently in existence and those yet to be made.

In the future, an examination of the quality of metadata available from Open Library, JSTOR, and other similar mass digitization projects is necessary. Larger scale studies, that consider additional metadata fields, are needed.

References

Ballard, T., & Lifshin, A. (1992) Prediction of OPAC spelling errors through a keyword inventory. *Information Technology and Libraries*, 11(2), 139-145.

Beall, J. (2005). Metadata and Data Quality Problems in the Digital Library. *Journal of Digital Information*, 6(3). Retrieved from http://journals.tdl.org/jodi/article/viewArticle/65

Chapman, A., & Massey, O. (2002). A catalogue quality audit tool. *Library and Information Research News* 26(82), 26-37.

Duguid, P. (2007) Inheritance and Loss? A Brief Survey of Google Books. *First Monday* 12(8). Retrieved from http://firstmonday.org/article/view/1972/1847

James, R. (2010). An Assessment of the legibility of Google Books. *Journal of Access Services*, 7(4), 223-228. doi: 10.1080/15367967.2010.503486

Nunberg, G. (2009a). Google Books: A Metadata Train Wreck. *Language Log.* Retrieved from http://languagelog.ldc.upenn.edu/nll/?p=1701

Nunberg, G. (2009b). Google's Book Search: A Disaster for Scholars. *The Chronicle of Higher Education*. Retrieved from http://chronicle.com/article/Googles-Book-Search-A/48245/

Oxford English Dictionary Online. Retrieved from http://www.oed.com

Pope,J., & Holley,R. (2011). Google Book search and metadata. *Cataloging & Classification Quarterly*, 49, 1-13. doi: 10.1080/01639374.2011.531234

Random.org. Retrieved from http://www.random.org

Ryans, C. (1978). A study of errors found in non-MARC cataloging in a machine-assisted system. Journal of Library Automation, 11(2), 125-132.

Townsend, R. (2007). Google Books: What's Not to Like? *AHA Today.* Retrieved from http://blog.historians.org/articles/204/google-books-whats-not-to-like