Alan Gevinson
Adjunct Professor
Department of History and Art History
George Mason University
Research focus: American Intellectual History

SUMMARY

Introduction: Scope and Method

In an effort to quantify current correctable problems with mass book-digitization projects from the point of view of scholars, I put together a list of titles relevant to the study of American intellectual history that included a significant number of titles published both before 1923 that are in the public domain and after 1922 for which copyright restrictions most likely apply, searched all titles in three mass book-digitization projects, and selected a more limited subset for a systematic examination of quality issues. Initially I searched 347 titles in Google Book Search (GBS), Microsoft Live Search Books (MLSB), and ACLS Humanities E-Book (HEB), and quantified the results of the search. For multivolume titles, I conducted searches in these projects for all volumes and quantified the results as well. The total number of volumes searched was 608. I selected 200 volumes for the systematic examination of quality issues and examined 80 digitizations in GBS, 80 in MLSB, and 40 in two projects directed more exclusively to the academic community: 39 in HEB and 1 title in Early American Imprints, Series I: Evans, 1639-1800 (EAI).[1]

Of the 200 volumes selected, 118 were published prior to 1923, while 82 were published after 1922. Of the 118 pre-1923 volumes, 62 were digitized by GBS, 55 by MLSB, and 1 by EAI. Of the 82 post-1922 volumes, 18 were digitized by GBS, 25 by MLSB, and 39 by HEB. For the pre-1923 volumes and the HEB post-1922 volumes, full-text digitizations were provided by the projects. For each of these volumes, I examined 100 pages in 5 sections of 20 pages each. Where possible, I selected the sections so that at least 50 pages separated each section. Because of rights restrictions, GBS and MLSB, with permissions, can provide only "previews" of post-1922 volumes, consisting of 10% or 20% of the total number of a volume's pages, or a more limited number of pages in close proximity to pages listed in search results. For post-1922 volumes, I selected the maximum number permitted, up to 100 pages, in 5 sections of near-equal numbers of pages, separated by at least 50 pages, including pages of notes, bibliographies, and indexes, whenever possible.

In the systematic examination of each book, I looked for

[1] EAI did not include a significant number of titles relevant to American intellectual history, so I limited my review within that project to one title (see entry 85 on p. 53 at http://www.clir.org/pubs/reports/pub147/data1Gevinson.pdf). Although problems noted in that entry cannot be assumed to be persistent throughout the project, I have included them in this report to demonstrate that more limited subscription digitization projects directed to the academic community, as well as mass-digitization projects such as GBS and MLSB, present significant specific problems.

problems with regard to accessing, downloading, and navigating; the physical condition of the books; the metadata offered; the quality of scans; and the accuracy of word searching. In Part I, I have indicated such problems for each title. In Part II, I have prepared a statistical summation report on recurring problems in each project.

In addition to an examination of digitizations pertaining to American intellectual history, I conducted two smaller studies. For Appendix A, I selected three titles of multilanguage reference works to survey specific problems relevant to this type of text. For Appendix B, I conducted a comparison analysis of 10 titles of scholarly volumes published prior to 1923 in any field that have been digitized in HEB and also in GBS, MLSB, or both. The number of digitizations examined in this analysis totaled 25. Appendix C provides bibliographic information in alphabetical order for the initial list of 347 titles relevant to American Intellectual History that were searched in GBS, MLSB, HEB, and EAI, and for the lists of titles in Appendixes A and B, in addition to notations indicating which projects included digitizations for each title and the type of digitizations (full-text, preview, snippets[2]) that were offered.[3]

Summary of Findings

Nearly one-third of GBS pre-1923 digitizations were of poor quality. Digitizations were considered to be poor when one or more of the following problems occurred: pages were missing; pages were unreadable; significant portions of pages were cut off or obscured by objects; pages appeared out of order; searching was not available for a significant number of pages; and significant numbers of pages were listed incorrectly in the page indicator, so that search results were compromised. Of the 200 volumes reviewed for quality issues, 24 digitizations were deemed to be of poor quality, or 12% of the total number. All but one were of volumes published before 1923, or 19.5% of pre-1923 volumes. Most were provided by GBS, which digitized 21 of the poor-quality digitizations, or 87.5% of all poor digitizations. 26.25% of all GBS digitizations checked were of poor quality. Of these 21 poor-quality digitizations, 20 were pre-1923 titles, or 32.3% of the number of GBS pre-1923 digitizations checked.

Inclusion in GBS and MLSB was dependent on the date of publication. Of the 347 American intellectual history titles searched, GBS

[2] For many post-1922 volumes for which GBS has not obtained permission to make available to the public its digitizations, GBS provides "snippets" of text in response to word searches—selected sentences in which search terms occur. Snippets can be of use to scholars because they can alert users to works in which relevant search terms occur. For many other post-1922 volumes, GBS offers only "no preview" listings, which have limited use for researchers, and because of this, have not been tabulated in this study.

[3] I conducted my initial examination of 200 volumes in January-February 2008. In March 2008, I checked a number of GBS digitizations previously reviewed and discovered that a number of problems had been corrected since the initial viewing. While I have revised my findings for those entries that I revisited in March, due to time constraints, I have not revisited every GBS entry, thus some of my critical comments may no longer be relevant.

provided full-text digitizations of 91% of volumes published prior to 1923; MLSB provided full-text digitizations of 61.5%. The numbers for inclusion dropped significantly in the subsequent period. GBS provided previews of 50.2% of the volumes published after 1922; MLSB provided previews of only 13.3%. With regard to specific time periods within the subset of post-1922 years, inclusion also was dependent to a large extent on date of publication. Of the volumes published between 1923 and 1950, GBS provided previews of 14%; MLSB provided previews of 2%. Of the volumes published between 1951 and 1980, GBS provided previews of 37.9%; MLSB provided previews of 3.4%. Of the volumes published after 1981, GBS provided previews of 74.3%; MLSB provided previews of 23.9%. These figures may be dependent on whether or not titles have remained in print.

Problems arose in searching for volumes. HEB provided the most varied ways for users to search for volumes, offering basic, Boolean, proximity, and bibliographic searches. GBS offered a sophisticated advanced search page, while MLSB allowed only basic searches of keywords or phrases placed within quotations. Because of this, MLSB search results often included irrelevant volumes.

Linkage to OPAC records can facilitate searching. All HEB volumes reviewed were linked to OPAC records in library catalogs. No OPAC links were found for MLSB digitizations, while 27.5% of GBS volumes checked were linked to OPAC records.

Searching for specific volumes within multivolume sets was problematic within all projects. When a search result does not indicate the specific volume number of a digitization, a common occurrence, users must open the volume and navigate to the title page in order to learn the volume number, a time-consuming process if the number of volumes in a set is large. When projects offer multiple digitizations of the same volumes of a large multivolume set with many editions, searching for a specific volume can take as long as 30 minutes.

Preview results of post-1922 volumes were more limiting in GBS than in MLSB. For post-1922 volumes for which GBS and MLSB provided "previews" consisting of 10% or 20% of the total number of a volume's pages, GBS offered consecutive pages at the beginning of books, but limited the number of consecutive pages viewable in the later portions. In contrast, MLSB lets users decide which 10% or 20% of a book to view, allowing the viewing of consecutive pages anywhere throughout the book. For a professor attempting to decide whether or not to use a book for a course, the MLSB model works better, as it allows the user to choose which specific pages to review.

Downloading options were more limited with GBS than with MLSB. Both GBS and MLSB provided options for downloading pre-1923 digitizations, while HEB, primarily a provider of post-1922 volumes, did not. MLSB provided downloading capabilities of interest to scholars that were not available in GBS. In downloaded MLSB

.pdf files, users can perform word or phrase searches, and can copy selected portions of text both as images from the page and as Optical Character Recognition (OCR) text, while in downloaded GBS .pdf files users cannot perform searches and can copy selected portions of text only as images from the page. In addition, the pages in downloaded GBS .pdf files are paginated in the page indicator according to the physical page number, rather than the number printed on the page image. In downloaded MLSB .pdf files, the page number in the page indicator designates the number printed on the page image, with the physical page number given in parentheses.

Navigation capabilities within volumes were more limited in MLSB than in GBS or HEB. Unlike in GBS or HEB, users of MLSB did not have the option to enter a page number into the page indicator to retrieve a specific page. MLSB provided only two ways to move forward and back to specific pages within digitized texts (in addition to clicking on contents links, many of which led to incorrect pages; see "other metadata issues" section below): one page at a time through clicking forward and backward buttons; or by moving a marker on a vertical ruler to the right of the page image, a device that forces users to rely on guesswork to determine the specific spot on the ruler that corresponds to a desired page.

Physical conditions accounted for OCR misreadings in more than one-half of pre-1923 volumes. Poor physical conditions within volumes—markings, folds, splotches, tears, stains, smudges, broken letters, and tape—were responsible for mistakes in OCR readings of 63.6% of the volumes published prior to 1923: 50% of GBS pre-1923 volumes, 78.2% of MLSB pre-1923 volumes. Physical conditions affected OCR readings adversely on 10 or more pages of 34.1% pre-1923 volumes reviewed: 11.3% of GBS pre-1923 volumes, 38.2% of MLSB volumes.

GBS and MLSB make limited use of MARC records and Library of Congress Subject Headings (LCSH). In its Book Search Help Center, GBS describes itself as "a book marketing program, not an online library."[4] Thus, GBS, like MLSB, does not offer MARC records, the source of authoritative bibliographic information that professional librarians have created to aid library collection control and to advance scholarship. Most important, MARC records offer subject headings derived from LCSH, the authoritative thesaurus that has proven to be extremely valuable in many fields of scholarship. With respect to many of the digitizations reviewed, especially for post-1922 volumes, GBS and MLSB have rejected the LCSH classification system available in MARC records and substituted their own headings, based on a system of classification that seems more relevant to marketing books than to promoting scholarship. In addition, for 37.1% of pre-1923 volumes, GBS provided links for the main subject heading from

[4] Google Book Search, "Authors: Common Questions." Available at http://books.google.com/googlebooks/author_faq.html.

MARC cataloging, but omitted subheadings. Clicking on the main subject links can lead scholars to thousands of entries listed in no understandable order, thus making such linkage useless. GBS does provide lists of "related books," but these listings are generated by "automated methods," according to GBS, and the standards employed to produce the lists are not offered to the public for examination.[5] In a number of cases discussed in Part II, so-called "Related books" bore no relation at all to subjects of volumes under consideration.

Although MLSB provided complete MARC subject headings for most pre-1923 volumes, MLSB limited its listings of subject headings by numbers of characters, often truncating headings by employing ellipses. Searching the headings omitted because of space did retrieve listings for the volume checked, and for other volumes to which the same heading had been assigned, but because these headings were omitted from the bibliographic information provided by MLSB, users must consult an outside source to find these headings in order to use them to find related texts. Listings of complete LCSH headings, as HEB provides for all volumes, would be of value for scholars interested in locating related texts.

Page number listings did not match MARC records (or the pagination within the volume itself, in the case of multivolume sets in which page numbers for individual volumes are not listed in MARC records) in 93.75% of the volumes digitized by MLSB and in 66.7% of post-1922 GBS volumes digitized. In these cases, GBS and MLSB used the total number of physical pages in books for their page number listings. MARC listings record the number of prefatory pages paginated in Roman numerals and offer notes regarding the inclusion of indexes and bibliographies, information of value to scholars.

Other metadata issues. In 40% of GBS volumes and 21.25% of MLSB volumes, contents links led to incorrect pages, listed incorrect or misspelled chapter titles, or were otherwise useless.

In 16.25% of GBS volumes, page indicator listings varied with the actual page numbers printed on page images. Because of this, scholars may have difficulty finding specific pages of text from citations located in other sources.

In its "about this book" reference page, GBS provides a host of links generated by automated methods: "key words and phrases" (terminology used for pre-1923 volumes) and "key terms" (terminology used for post-1922 volumes); "references from books"; "popular passages"; and "related books." An examination of these links revealed that they are of limited use for scholars. Listings of "key terms" often exclude terms that are more "key" than those included; links for "key terms" often do not lead to all pages in which terms are discussed, and on occasion the links are totally useless, so that pursuing them is a waste of time. GBS offered no rationale for the selection of books included in "references from books." On occasion, links in this section did not lead to relevant pages. Some links in the

[5] Google Book Search, "Google Book Search Help Center." Available at http://books.google.com/support/bin/answer.py?answer=53549.

"references from scholarly works" section led to books that bore no relation to the book in question. Often the quoted passages from the "popular passages" section were from texts that the author had quoted, not the author's own words. Often page numbers were incorrect, as were the number of books listed that purported to include the quote. Some of the books listed in the "related books" section had little or no apparent connection to the text under consideration.

Quality of scans. In 21% of pre-1923 GBS volumes, one or more pages were missing. In 9.7% of pre-1923 GBS titles, one or more pages were illegible. In 19.4% of pre-1923 GBS volumes, portions of one or more pages were cut off or obscured. In 16.25% of pre-1923 GBS volumes and in 12.7% of MLSB pre-1923 volumes, OCR readings were adversely affected by light, dark, or blurry pages; portions of pages; or individual letters.

In general, the quality of the scanned image was sharper in MLSB than in GBS or HEB.

Word searching within texts. GBS retrieves a maximum of 30 hits of pages in which search terms occur, unlike MLSB and HEB, which have no limits on numbers of pages retrieved. GBS does not indicate how it selects the 30 hits to display when search terms occur on more than 30 pages.

Only HEB displays retrieval results of each hit occurrence per page. GBS and MLSB display results of only one hit occurrence per page.

Only HEB allows users to truncate search terms, so that retrieved results for the same search can include plural and singular forms of words, possessive forms of names with nonpossessive forms, and other related forms. For example, "federalist*" retrieves "federalists" and "federalist"; "Lazarsfeld*" retrieves "Lazarsfeld's" and "Lazarsfeld"; "pragmatis*" retrieves "pragmatism," "pragmatist," and "pragmatists."

MLSB purports to sort search results within texts by "relevance," while GLS and HEB sort results by the order in which hits occur in texts. In fact, MLSB orders results through automated means that can make it hard for scholars to find the most relevant listings when searches retrieve large numbers of hits.

Mistakes in OCR readings occurred in 88.5% of the volumes reviewed. In addition to physical conditions and poor scanning, OCR misreadings occurred most often in footnotes, in italicized words, in foreign-language words, in words with letters in all caps, in words beginning with a capital letter, in words hyphenated between lines, in index pages, and in dates.

In HEB, OCR misread the numeral "1" as the letter "I" in 48.9% of all HEB volumes checked. Searching for dates was compromised as a result of these errors. In 20.5% of HEB volumes, searching was inaccurate because OCR read separate words as one word.

In all of the projects, OCR misread many words that were hyphenated between lines as two separate words. None of the projects

retrieved any words or phrases (including names) that were hyphenated between or extended over two pages.

Examination of multilanguage reference works (Appendix A). The examination of multilanguage reference works highlighted a problem that appeared occasionally in the review of American intellectual history volumes. The accuracy rate of OCR reading of words in languages other than English was much lower than that of English-language words. The accuracy rate with regard to words in Latin was especially low.

Comparison examination of 10 pre-1923 works (Appendix B). The comparison examination unearthed a problem in HEB that did not occur with respect to the American intellectual history volumes reviewed, and it highlighted another problem in HEB. For every page in two of the volumes examined, HEB placed incorrect page numbers in page indicators. As a result, retrieval of results from searches was difficult. While results lists presented correct page numbers for pages in which search terms occurred, clicking on the links for these page numbers retrieved pages for which the numbers appeared in the page indicators, which were not the pages in which the search terms occurred. In addition, in one text, OCR in HEB read footnote indicators as parts of words, which also compromised searches.

In one HEB text, the numeral "1" was read by OCR as "I," a recurring problem in HEB that made searching of dates unreliable. Like GBS and MLSB, HEB presented a high percentage of digitizations in which OCR misreadings occurred on three or more pages. Otherwise, HEB did not present major problems.

GBS and MLSB presented problems consistent with those analyzed in the review of American intellectual history volumes. In 75% of the GBS entries, GBS provided subject headings that used LCSH classifications without the subheadings provided by MARC records. This made many of the headings useless as links to related books because the broad terms linked (for example, "United States") retrieved too many digitizations—ordered in no understandable manner—that were only marginally related to the texts at hand. In 71.4% of MLSB volumes checked, poor scanning led to missing pages, illegible pages, pages not read by OCR, or pages with portions cut off or obscured. MLSB consistently listed the number of physical pages in volumes rather than the number of pages listed in MARC records. Both GBS and MLSB included a significant number of bad contents links. In both GBS and MLSB, the percentage of volumes in which OCR did not read text correctly in indexes, footnotes, or bibliographies was high: 50% in GBS and 85.7% in MLSB.

The comparison examination brought to light an additional problem common to all projects: margin headings often were read by OCR as part of the line of text that they followed. As this problem did not occur with every margin heading that appeared in texts, it is conceivable that projects can employ methods to correct it. The comparison examination also highlighted a high inaccuracy rate in OCR readings of texts in Old English.

Recommendations

1. Many errors that have led to poor-quality evaluations of GBS digitizations—missing pages, unreadable pages, obscured or cut-off pages, pages out of order, pages without OCR readings, pages listed incorrectly in the page indicator—could have been prevented through better scanning and quality control operations. Digitizations of books with physical conditions that tend to cause OCR misreadings—markings within text (especially in ink), tears, smudges, etc.—should be carefully checked and corrected before public display. Areas of text with a high incidence of OCR misreadings—footnotes, italicized words, index pages, etc.—should be scanned more carefully than other areas. Projects should pay special attention to find methods to resolve scanning problems that occur at a high frequency.
2. When projects cannot avoid poor-quality digitizations, they should label them as such and direct users to better digitizations of the same text if they exist. Projects should ensure that the first digitization listed in a search results page is the best-quality digitization for that project. Missing pages should be identified in metadata and within texts.
3. As some 86% of books surveyed in this study that were published between 1923 and 1950 were not available for viewing or searching, efforts should be made to remedy that situation, whether through a concerted approach to the publishing community or to Congress to change the copyright laws in order to better "promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."[6]
4. GBS and HEB should employ better scanning technology so that the visual quality of page images is as good as that delivered by MLSB.
5. All projects should provide users with the ability to conduct advanced searches for materials and the ability to order results of searches for books according to title, author, publication date ascending, and publication date descending, as in the Library of Congress online catalog and university library OPACs.
6. Volume numbers of multivolume works should be indicated in search retrieval results.
7. MARC-derived metadata should be provided by all projects. Marketing-oriented subject classifications and machine-generated "key terms" and "related books" should not replace tried-and-true tools for scholarly subject access, such as subject headings derived from LCSH. The omission of LCSH subheadings, as practiced to a large extent by GBS, should cease, as the resultant subject heading links often are too broad to be useful. GBS must employ professional catalogers or subject specialists to augment metadata derived from automated methods before scholars will trust the reliability of these devices for serious research purposes.

[6] U.S. Constitution, art. 1, sec. 8.

8. Projects should alert users that searching often does not retrieve words hyphenated between lines and never retrieves words hyphenated between pages, unless projects can correct these recurring problems. Users should be informed that word searching, especially for names, should not replace consultation of book indexes.
9. Truncation for word or phrase searching should be incorporated into GBS and MLSB.
10. MLSB should sort the results of word searches by the order in which hits appear within texts, rather than order results by a so-called relevance achieved through automated methods. Authors often develop their discussions progressively through their narratives. Scholars may want to follow the progressions that authors have employed in order to fully appreciate the subtleties of their discussions. MLSB's system of sorting by relevance makes it difficult to do this.
11. In the interests of scholarship, it is important that access to non-subscription mass digitization projects always remain open to all users. I was unable to access MLSB on my home computer for a number of weeks, a condition that I found was not unique to me. As with other problems encountered during this examination, the access problem resulted in wasted time.

Conclusion

In a 1989 hearing of a Senate subcommittee presided over by Senator Al Gore that was convened to create legislation for research and development of "a national network of information superhighways," Librarian of Congress James H. Billington portrayed the Library's collections as "the freight that can be carried on this highway." In a prepared statement to the subcommittee, Billington related,

> With 88 million items in the Library, we have the largest collection of recorded information and knowledge ever assembled in one place here on Capitol Hill. The Library of Congress represents the nation's most important single resource for the information age. The proposed establishment of a National Research and Education Network would give an immense boost to the access of this material and allow the Library of Congress to provide to the country much more of its unequalled data and resources which can now be obtained only by visiting Washington.[7]

Nearly two decades later, despite the efforts of the mass-digitization projects surveyed in this study, the world of scholarship remains in need of the envisioned National Research and Education Network or some other type of endeavor to allow scholars and other users

[7] Congress. Senate. Committee on Commerce, Science, and Transportation. *National High-Performance Computer Technology Act of 1989: Hearings before the Subcommittee on Science, Technology, and Space of the Committee on Commerce, Science, and Transportation*, 101st Cong., 1st sess., September 15, 1989.

open, reliable, and comprehensive access to the "unequalled data and resources" available in the world's great repositories of information and knowledge.

Google proclaims that its "mission is to organize the world's information and to make it universally accessible and useful."[8] As noted above, GBS also identifies itself as "a book marketing program, not an online library."[9] Libraries long have taken the lead in organizing information to aid and foster scholarship. Replacing the transparent methods and standards of library professionals with automated methods and marketing-oriented devices based on algorithms not available to the public or subject to review by outsiders has resulted in digitizations of more limited usefulness to scholars than might have occurred otherwise. The world of scholarship would benefit more from online libraries than book marketing programs, and from digitization projects of better quality and with more transparent methods and standards than GBS and MLSB.

[8] Google, "Corporate Information. Company Overview." Available at http://www.google.com/corporate/.

[9] Google Book Search, "Authors: Common Questions." Available at http://books.google.com/googlebooks/author_faq.html.