Google Books as a General Research Collection

By Edgar Jones

The current study attempts to measure the extent to which "full view" volumes contained in Google Books constitute a viable generic research collection for works in the public domain, using as a reference collection the catalog of a major nineteenth-century research library and using as control collections—against which the reference catalog also would be searched—the online catalogs of two other major research libraries: one that was actively collecting during the same period and one that began actively collecting at a later date. A random sample of 398 entries was drawn from the Catalogue of the Library of the Boston Athenæum, 1807–1871, *and searched against Google Books and the online catalogs of the two control collections to determine whether Google Books constituted such a viable general research collection.*

> "There's an east wind coming, Watson."
>
> "I think not, Holmes. It is very warm."
>
> "Good old Watson! You are the one fixed point in a changing age. There's an east wind coming all the same, such a wind as never blew on England yet. It will be cold and bitter, Watson, and a good many of us may wither before its blast. But it's God's own wind none the less, and a cleaner, better, stronger land will lie in the sunshine when the storm has cleared."
>
> —Arthur Conan Doyle, *His Last Bow*

On December 14, 2004, Google announced that it had concluded agreements with five major research libraries to begin what is now known as the Google Books Library Project.[1] The libraries—the so-called Google 5—were the New York Public Library and the libraries of Harvard, Michigan, Oxford, and Stanford universities. These libraries agreed to let Google digitize volumes from their printed book and serial collections in exchange for institutional copies of the digitized volumes.[2] While the agreements set broad parameters for cooperation, Google gave the libraries sole discretion in determining the volumes to be digitized.

The Library Project—and the discretion given the libraries in determining which volumes would be digitized—raises an interesting question: To what extent is Google creating a research collection? Coyle has suggested that the manner in which collections are being selected for inclusion in the Library Project—many being taken *en bloc* from low-use remote storage facilities—makes it difficult to characterize Google Books as a "collection" in the accepted sense, though for better or worse "it will become a de facto collection because people will begin using it for research."[3] Is this true? Is this testable? Can sheer volume, in fact, render

Edgar Jones (ejones@nu.edu) is Bibliographic and Metadata Services Coordinator, National University Library, San Diego, California.

Submitted July 7, 2009; tentatively accepted September 17, pending modest revision; revised and resubmitted September 22 and accepted for publication.

moot the role of selection in this case? The current study attempted to answer these questions.

While the focus of this study was on content digitized by Google through 2008, one should keep in mind that the volume of available digitized content continues to grow. Since the initial Google 5 cooperative agreements at the end of 2004, Google has entered into agreements with an increasing number of research libraries, both in the United States and abroad, while the European Union has begun funding a digitization program of its own centered on the collections of European cultural heritage institutions (libraries, archives, and museums).[4] Initially, there also was competition from elsewhere in the commercial arena, but this proved to be comparatively short-lived. Within a year of the Google announcement, Microsoft, in cooperation with the Internet Archive, began to digitize print content from several libraries under the rubric of Live Search Books. In May 2008 this effort was abandoned, though content already digitized under that program—some 750,000 volumes—remained available via the Internet Archive.[5]

In terms of scope, several of the Library Project partnerships cover both older public domain materials and more recent publications still subject to copyright protection. To this extent they complement Google's partnerships with publishers to provide access to a continuity of content across time periods.

This continuity of content is important from Google's perspective. In Google's December 2004 press release, cofounder Larry Page set the Library Project in the context of his firm's stated mission "to organize the world's information and make it universally accessible and useful."[6] As a search engine, Google's principle interest in digitizing printed materials is in indexing the content, both structured and unstructured, to enhance search results. In its business model, Google uses search terms and results as triggers for the online display of related advertising. By providing additional indexed content from Google Books (and the Library Project), Google both increases the usefulness of its flagship search engine (by incorporating results from Google Books as well as other sources) and makes it more appealing to advertisers (by increasing the potential customer base to include researchers and other interested parties).

As has been noted frequently, Google is digitizing on an industrial scale, indeed on a scale unlike anything seen before.[7] The process is easy to describe. Books are removed from the shelves, barcodes are scanned—to change the volume's circulation status and to extract the related metadata from the catalog—and the volumes are removed to a Google facility for digitization. Google digitizes the individual page, subjects the digitized images to sophisticated (if not foolproof) optical character recognition (OCR), and finally indexes the OCR–extracted text. The digitized page images may be freely available for public viewing (if determined to be in the public domain), or viewing may be restricted in some way, depending on the copyright status of the digitized work (or one or more of its components) and the nature of Google's agreements with the publisher.[8]

What Is a Research Collection?

While other digitization programs on various scales also are under way (as noted above), none approaches the scale or ambition (or potential for market dominance) of Google Books. For this reason the volumes digitized by Google seemed the most appropriate objects of which to ask: Do these digitized volumes in themselves now constitute a viable general research collection? This may seem a fairly straightforward question, but it raises an antecedent question: What is a research collection?

In the abstract, a research collection is a collection of materials used primarily to support research (as opposed to one that supports teaching and learning or one that is used primarily for recreational purposes). Unfortunately, this definition does not lend itself to objective measurement, and it says nothing about the content of such a collection, since, in theory, *any* collection can support research of *some* kind.

Indeed, most research collections are developed to address the needs of a particular research community, a community that will typically reflect a variety of research interests and intensities. This variety will itself change over time. Research collections are by their nature complex. Such complexity underpins the design of the Conspectus model developed in the early 1980s by the Research Libraries Group for cooperative collection development. The Conspectus asked participants to characterize their collections according to a variety of parameters, including research area (defined by ranges within a bibliographic classification scheme or by subject descriptors), language, geographical scope, chronological periods, formats, and collection depth (this on a five-point scale, with 4 indicating "research level").[9] This produced a nice matrix for describing the variety of possible research collections, but it also made clear that the idea of a "generic" research collection was an oxymoron.

Ideally, in addition to being designed for a particular research community, a research collection also satisfies the needs of that community. But while research collections can come in a variety of shapes and sizes, data suggest that whatever their shape and size, local researchers will always feel that their own library's collection falls short—this despite years of earnest collection development by librarians. At member institutions of the Association of Research Libraries, for example, respondents to successive LibQual+ library service quality surveys routinely report that their libraries provide inadequate support for their research needs. On three LibQual+ items measuring collection support for research, the "adequacy gap"—the degree to which an item exceeds (or not) a user's minimum requirements—has typically been a negative number.[10] To put as generous a spin as possible on the meaning of these LibQual+ responses, one could say that research collections always must be works in progress.

Although there may be no such thing as a typical research collection, at major research universities—where the research communities are larger and more multidisciplinary—a certain amount of homogeneity can be expected to develop across the associated research collections. It is not unreasonable then to treat one of these collections as approximating a "generic" large university research collection.

Having posited that the collections of large research libraries approximate to a generic research collection, a question remains about how much unique content is found in a typical research library. No one knows for sure, but overlap studies suggest it is more than one might expect.[11] In a 2005 study (shortly after the announcement of the Google Books Library Project) Lavoie, Connaway, and Dempsey determined that the Google 5 libraries collectively held about one-third of the resources cataloged as books in the OCLC WorldCat database, and most of these resources—61 percent—were unique to just one of the five. This percentage of uniquely held resources increased with the age of the resources involved, with 74 percent of resources published between 1801 and 1825 being held by just one Google 5 library.[12] However, there is unique and then there is *unique*. In subsequent research, Lavoie and Schonfeld examined a random sample of one hundred WorldCat records for such "uniquely held" resources and found that "many of the English language materials appear to be locally-produced ephemera rather than traditional published books."[13] This suggests that unique holdings may be less of a problem for the idea of a "generic" research collection than might otherwise appear.

Given its existing digitization agreements, it seems reasonable to expect that Google Books will eventually become a generic research collection in this sense. It will at the very least become the University of Michigan Library, since Michigan has agreed to digitize its entire library (aside from Special Collections).[14]

Scope of the Study: The Public Domain

This study is limited to materials that would have been viewable in full in Google Books in 2008 under current copyright law, that is, materials then in the public domain. This was a pragmatic decision in that such materials are the only ones that can be demonstrably shown to have been digitized. But it also attempted to address the idea of the research collection as implied in Coyle's original comment—something that would be used for research.[15]

The author also felt that restricting the study to public domain materials would produce a more conservative estimate of overlap than would be true for more recent imprints, given the participation of publishers as well as libraries for this material in Google Books, the growth and expansion of American research libraries over the last century, and Lavoie, Connaway, and Dempsey's findings of increasing overlap within WorldCat as imprints become more recent.[16] This suggested that whatever conclusions the study reached about materials in the public domain would apply with even more force to materials still covered by copyright should those materials be exposed to viewing via Google Books at some future point.

How big is the public domain? Nobody knows. The online English Short Title Catalogue, covering pre–1801 materials published chiefly in the British Empire and the United States, had accumulated nearly half a million entries by mid-2009.[17] More generally, Buringh and Van Zanden estimated that more than 1,750,000 books (defined as fifty pages or more) were published in Western Europe prior to 1801.[18] To bring this closer to the present, in a 2006 examination of books in OCLC WorldCat, Lavoie and Schonfeld found that roughly six million books (18 percent of the books then in WorldCat) were published prior to 1923 (a rough indicator of public domain status in 2008).[19] The chart accompanying the Lavoie and Schonfeld article suggested a steady rate of increase in publishing output throughout the nineteenth century.

But as anyone who has tried to establish the copyright status of a book knows, public domain status is not simply a question of date of publication, and Google's caution increases as the date of publication approaches 1923 (with digitized volumes more likely to offer "no preview," "snippet," or "limited preview" access rather than "full view"). For that reason, the current study adopted a more conservative interpretation of public domain, looking for a cutoff farther back in the nineteenth century.

Given this working definition of the public domain, one question that still needs to be addressed is that of page image quality. Scholars have not been silent on this question, and anecdotal evidence suggests the frequency of poor page image quality is not insignificant, though the solution is by no means clear.[20] Quality control of scanned texts, unlike the initial scanning, cannot be done on an industrial scale and is labor-intensive. In theory, however, volumes to be rescanned could be prioritized on the basis of user complaints. A "flag this page" feature, presumably for this purpose, was present in earlier versions of Google Books but had been discontinued at the time of writing. Similarly, the problem of nontextual content (illustrations, maps, etc.) in the digitized volumes, while not significant for indexing, is significant for research, especially in the case of folded material, which Google scanners typically digitize in its folded form.[21]

Given both this more restricted definition of the public domain and the caveats regarding image quality and folded materials, the current study attempted to measure what proportion of the volumes in a generic research collection (as described above) would be present in digitized form in Google Books, using as a reference collection the catalog of a major research library that was actively collecting during the period, and using as control collections—against which the reference catalog also would be searched—the online catalogs of two other major research libraries, one actively collecting during the same time and one that began actively collecting at a later date.

The general research question can be stated formally as follows: Given that A and B were major American research libraries during the nineteenth century, while C became one during the same period, is there a greater probability of manifestation-level overlap between A and Google Books, between A and B, or between A and C? The question can be answered by testing two hypotheses:

1. A larger proportion of a random sample of entries drawn from A's catalog will be found in Google Books than will be found in the online catalog of B.
2. A larger proportion of a random sample of entries drawn from A's catalog will be found in Google Books than will be found in the online catalog of C.

Experimental Design

The experiment involved six steps:

1. identifying an appropriate reference collection
2. extracting a random sample of entries from the reference collection
3. searching the metadata from those entries against Google Books for matching manifestations
4. identifying appropriate control collections
5. searching the metadata from the reference collection entries against the OPACs of the control collections
6. recording the results and performing appropriate statistical tests

Identifying the Reference Collection

What would serve as an acceptable reference collection? More precisely, what contemporary collection would meet the criteria for a major American research collection? Only the largest libraries of the nineteenth century might lay claim to such a distinction, and of these it would be necessary to select one with a reliable printed catalog to efficiently select a random sample of entries for testing.

Identifying the major American research libraries of the nineteenth century is not difficult, but reliable comparative statistics are rare. As noted in the landmark *Public Libraries in the United States*, for much of the nineteenth century library statistics were collected and published only occasionally, and reports were often of dubious reliability.[22] In his 1851 survey, Charles Jewett, then librarian of the Smithsonian Institution, listed five libraries as having collections of at least 50,000 volumes: Harvard University, Yale College, the Philadelphia Library (including the Loganian Library), the Library of Congress, and the Boston Athenæum. Harvard's libraries collectively held the largest number of books at the time: 84,200 volumes.[23] Less than two decades later the largest libraries had grown—Justin Winsor's survey in 1868–69 reported that there were now six libraries containing at least 100,000 volumes—but the ranks had changed. Half of the six largest were newcomers (Winsor's own Boston Public Library and the Astor and Mercantile libraries in New York City). Only the Library of Congress, Harvard University, and the Boston Athenæum continued from Jewett's earlier list.[24] And of these three, only the Boston Athenæum had a contemporary published catalog that could serve as a reliable data source for the current study: the five-volume catalog compiled under the direction of Charles Ammi Cutter and published between 1874 and 1882.[25]

The use of any library catalog as a surrogate for a generic authoritative research library will be biased to the extent that it necessarily reflects the collection interests of the library involved. In the case of the Boston Athenæum, this will be a bias toward topics of interest to its members and toward printed books available since the library was founded in 1807 (leading to an underrepresentation both of non-American imprints and of older books), as well as a slight regional bias. However, against this unavoidable bias must be set the clear intention of the founders to create an exceptional general research library "containing the great works of learning and science in all languages, particularly such rare and expensive publications as are not generally to be obtained in this country."[26] If one accepts the judgment of contemporaries, then the founders achieved their purpose. According to Jewett, the Athenæum library was "hardly surpassed either in size or in value by any other in the country."[27] According to Edwards, it "[stood] saliently out from amongst its compeers alike for its extent, its liberality of access, [and] its richness in departments not usually well-filled in American Libraries."[28] So while taking the Boston Athenæum as representative of a generic major research library in some absolute sense may not be possible, it can be described with confidence as one of the major American research libraries of that time.

Cutter's catalog was said to include 92,000 volumes and 36,000 pamphlets.[29] It does not include works from the fifty public domain years that followed its publication, and given the increase in research collection size and overlap during this period observed by Lavoie et al., the use of Cutter's catalog in the current study will tend to underestimate the actual degree of overlap between the reference collection and Google Books, i.e., the reported (pre–1872) hit rate in Google Books will likely be less than it would be for the entire (pre–1924) "public domain" period.

Extracting a Random Sample from the Reference Collection

The objects of interest were manifestations represented by main entries in the *Catalogue of the Library of the Boston Athenæum, 1807–1871.*[30] While Cutter did not go so far as Panizzi, who included his cataloging rules in the first volume of his aborted 1841 *Catalogue of Printed Books in the British Museum,* he did include one page of detailed explanations of the organization and structure of entries in the catalog, including the determination of the main entry.[31] Cutter entered works of personal authorship under the name of the author, collections under the name of the editor, and works of corporate responsibility (interpreted broadly) either under the name of the body, territorial authority, or—for local institutions—under the name of the place where it was located.

The five volumes of Cutter's catalog are numbered consecutively, and the author of this study used a random number generator to create the sample of pages from which main entries would be extracted. For each page in the sample, the first main entry appearing on the page was selected. In cases where no main entry appeared on a page (for example, the page consisted entirely of subject entries), the first main entry appearing on a subsequent page was selected. While Cutter made entries for the component parts of aggregate works, these were ignored in selecting the sample unless they represented physically distinct volumes (rather than contributions to a single volume) to restrict the sample to physically substantial works. Similarly, physically separate volumes of fewer than fifty pages were omitted from the sample.

Searching the Sample in Google Books

The author used the metadata from the sample entries in Cutter's catalog—specifically the author (when present) and title—to search Google Books for corresponding entries. In cases where a corresponding entry was not found in Google Books, the Boston Athenæum's online public access catalog Athena (http://catalog.bostonathenaeum.org) was searched on the chance that more complete metadata might be located. If no corresponding entry was found in Athena, the lists of corrections at the beginning of each of the first four volumes of Cutter's catalog (and the end of the fifth) were searched for unsuspected typos and similar errors.

In cases of aggregate works—multivolume monographs and serials—the author deemed a manifestation to be a match if digitized images of at least 50 percent of its component physical volumes were present in Google Books; otherwise it was deemed to be absent from Google Books.

In terms of Functional Requirements for Bibliographic Records (FRBR) Group 1 entities, Google Books is technically not a collection of manifestations but rather of images of items (specific exemplars) that have been digitized from participating library collections.[30] This anomaly is recognized in the FRBR discussion of reproductions, but it may have practical consequences in that a given alteration to a specific item may add or subtract value from a given manifestation, depending on whether the alterations are positive in form (e.g., scholarly annotation) or negative (e.g., vandalism).[33]

While in many cases a given manifestation will not be represented in Google Books in digitized form, in other cases it will be represented multiple times by exemplars from different institutions (and sometimes from the same institution). This is an unavoidable consequence of digitization undertaken on such a scale and on such an accelerated timetable. The current study did not attempt to count such instances of multiple digitized items, though instances were not uncommon.

The author considered a digitized volume in Google Books a match if it included all elements of the manifestation as described in Cutter's catalog (augmented when necessary by elements from the description of the same manifestation in the Boston Athenæum's Athena online catalog or other sources). Digitized volumes that appeared to represent the same text (what FRBR calls an expression) were rejected if they differed on any element of bibliographic description; that is, identity of container took precedence over identity of content. The author felt this was necessary to ensure rigorous matching criteria, since differences in text from one manifestation to another—signaling a different expression—are not always readily apparent from bibliographic data. One consequence was that legislative documents represented singly or in small groups in Cutter's catalog could not be linked to the bound volumes that superseded them in the collections of the Google Books partner libraries.

Identifying the Control Collections

The author also used the metadata from the sample entries in Cutter's catalog to search entries in the OPACs of two major research libraries (defined as Association of Research Libraries member institutions with reported holdings greater than eight million volumes).[34] One of these selected institutions was actively collecting during the same period as the Boston Athenæum, while the other began collecting at a later date.

Searching the Sample in the Control Collections

In searching the OPACs of the control collections, the author defined a match as a record representing the same manifestation or a reproduction of that manifestation as defined in *FRBR.*[35] The decision to include reproductions as matches was made on the basis that Google Books manifestations are de facto reproductions, and reproductions in the control collections should therefore be treated as equivalent. In some cases, fidelity to the original might not be perfect, particularly in the case of originals that included color versus a reproduction made in monochrome (typically a microform copy or a digitized copy made from a microform copy).

In many cases with older books, research libraries have purchased collections of books listed in a given printed bibliography (e.g., Charles Evans' *American Bibliography* (1639–1820), originally supplied by Readex Microprint Corporation on micro-opaque, then on microfiches, and now online).[36] In the current study, items in such collections were "matches" only if they were represented by records in the library's OPAC. This inevitably resulted in undercounting, especially in the case of the control library that began its collecting during a later period than the Boston Athenæum. The author made the assumption in this case that a resource that was not represented by a record in the OPAC would be invisible to most library users seeking it and perceived as not being held by the library concerned. For a user to track down a desired resource in such cases would require a detailed knowledge of bibliography, the identity of major microform and digital collections, and the highways and byways of the website of the library concerned—a knowledge far beyond that of most library users.

As with Google Books, the author considered a manifestation of aggregate works a match if digitized images of at least 50 percent of its component bibliographic volumes were present in a control collection as determined from its OPAC; otherwise it was deemed to be absent from the control collection. In those rare instances where the extent of holdings of an aggregate work could not be determined from data in the library's OPAC, it was deemed to be a match for the purposes of the study. This decision was based on an extrapolation of the pattern for other aggregate works in the sample where the control libraries typically owned more than half of the bibliographic volumes concerned. In cases where sample entries represented component parts of aggregate works (e.g., an analyzable component of a series), the author searched the OPAC under indexed elements (creator, title, and so on) for both the component work and the aggregate work.

Performing the Statistical Tests

Once the author searched Google Books and the OPACs of the control institutions for the same sample entries, estimates were made of the proportion of manifestations represented by entries in Cutter's catalog that were likely to be found in (a) Google Books, (b) the collection of the older control institution, and (c) the collection of the later control institution. Z-tests were then performed to determine whether the proportion of entries found in Google Books was significantly different from the proportion found in either of the control collections.[37]

The Process

Problems encountered during the current study can be subdivided into problems determining the extent of the manifestation represented by an entry in Cutter's catalog, identifying a manifestation in Cutter's catalog, locating a manifestation in Google Books, and identifying manifestations in the catalog of the control collections. These are discussed below.

Extent of an Item

Cutter's catalog includes entries for both books and pamphlets. The inclusion of pamphlets makes it an unusually thorough catalog. Unfortunately, it also makes it too unusual to justify including pamphlets in the current study. Because of their brevity and their often ephemeral nature, pamphlets are seldom cataloged at the level of the individual pamphlet, so identifying their presence or absence unequivocally in Google Books or the control collections would have been problematic.

To exclude pamphlets on a consistent basis, the study adopted a modified version of the UNESCO definition of "book": "a non-periodical printed publication of at least 49 pages, exclusive of cover pages . . . made available to the public."[38] For purposes of the study, the author modified the UNESCO definition to include periodicals, even when shorter than forty-nine pages. Application of the definition in practice was complicated by the fact that Cutter seldom recorded the extent of an item in his catalog (other than for multipart items). Overcoming this obstacle required determining an item's extent by searching Athena, the online catalog of the Boston Athenæum, or occasionally the online catalog of the Countway Library of Medicine at Harvard University (where the library of the Boston Medical Society—formerly in the Boston Athenæum—currently resides) for a fuller bibliographic record. Metadata in Google Books turned out to be unreliable in this regard, as different Google Books metadata purporting to describe the same manifestation occasionally presented contradictory data as to the manifestation's extent.

Identifying a Manifestation in Cutter's Catalog

One difficulty encountered when examining entries in Cutter's catalog resulted from the extensive manipulation of title data by the cataloger preparing the entry. This manipulation was necessary to take full advantage of the different environment constituted by a closed system such as Cutter's book catalog (compared to the open systems of today). In a book catalog, entries are interpreted in the context of the surrounding entries, and in the best such catalogs they are crafted to exploit that context to the utmost. Cutter's catalog reduced titles to the extent that they conveyed the maximum amount of information with a minimum amount of text, eliminating text that was redundant with the introductory heading and also text that added nothing of substance to the title. For example, Cutter reduced "Catalogue de la riche bibliothèque de D. José Maria Andrade" to "Andrade, José Maria. Catalogue de [sa] bibliothèque," and he reduced "Report of the debates in the Convention of California on the Formation of the State Constitution, in September and October, 1849" to "California. Constitutional Convention, 1849. Report of debates."

Cutter assumed the users of books in foreign languages would be familiar with those languages, an assumption that gave him greater latitude in his manipulation of the entries but that presents some problems to the modern researcher. This is especially the case with titles in Greek or Latin, where the elimination of introductory words or phrases can often alter the grammatical cases of the remaining title words. Being able to reconstruct the title as it appears on the item facilitates searching in Google Books (always providing the title has not been similarly, or differently, abridged in the metadata supplied to Google).

In rare instances, Cutter's catalog altered a title to make it more descriptive of a book's content. For example, Edward T. Channing's *Lectures Read to the Seniors in Harvard College* was rendered by Cutter as "Lectures on rhetoric and oratory; biographical notice by R. H. Dana." Such cases presented particular challenges when searching Google Books or the OPACs of the control collections.

Like all catalogs, Cutter's catalog was of course subject to error. Cutter's staff were not immune to the odd typographical or other error, and Cutter himself did not catch all of these (as witness the multiple pages of corrections appended to each volume after the first). Whenever possible, the author verified entries in Cutter's catalog against bibliographic records in Athena. In rare cases, the volumes listed were no longer owned by the Athenæum.

Locating a Manifestation in Google Books

There was no problem in identifying the sample manifestations once they were located because Google Books contains digitized page images, including the title pages. There were, however, problems *locating* the sample manifestations.

If the title was not distinctive, the search result might include both volumes from the target publication and volumes from other publications with the same title mixed together randomly. For example, clicking on the "other editions" link from "Report of the Treasurer" in Google Books produced a list that included reports from the treasurers of Alabama, Connecticut, Maine, and several other states. This was somewhat ameliorated by the fact that Google Books initially sorts "other editions" by date of publication.

If the title was subject to varying treatment ("cataloger judgment") by the originating catalogers or by the cataloging rules in force at the time the item was cataloged locally, the same manifestation might be scattered across the pages of a Google Books search result. For example, of the four digitized sets of Cutter's Boston Athenæum catalog in Google Books at the time of the current study, the metadata supplied by Harvard and New York Public Library used the full title from the title page (*Catalogue of the Library of the Boston Athenæum: 1807–1871*) while the Michigan metadata omitted the terminal dates and the Oxford metadata (taken from the Bodleian pre–1920 catalog) followed older cataloging practice in omitting the text that was redundant with the principal access point by reducing the title to "*Catalogue . . . 1807–1871*." Had the Oxford copy been the only one digitized, locating the manifestation in Google Books would have been problematic. When searched in early November 2008 using the "Advanced Book Search" page to limit by viewability and dates of publication, the Oxford copy appeared on page seven of ten in the result set.

To be weighed against this confusion is the serendipity resulting from the acts of digitization and indexing. In one case, a particular manifestation was discovered at the end of a digitized volume, bound with a different manifestation: Oxford University's copy of Henri Storch's *Considérations sur la nature du revenu national* appears after page 428 of volume 4 of his *Exposition des principes qui déterminent la prosperité des nations*. It was located not by searching against the Google Books metadata but by searching against the full text. It was not listed in Oxford's OPAC, suggesting that in this case the Google digitization may have turned up a hidden work in Oxford's collection.

Similarly, Google Books may serendipitously discover works that are significant at the item level. In one case from the sample, an item's provenance could be traced from the author to then-Vice President John Adams to the American Academy of Arts and Sciences and finally to Harvard University, with some minor annotations in Adams' hand. The provenance was unremarked in HOLLIS (http://discovery.lib.harvard.edu), the Harvard online catalog, where the item was in the general circulating collection.

The author used the metadata in Google Books sparingly, since its reliability was open to question. Limiting a search by certain metadata elements occasionally produced anomalous results. For example, a search for "full view" books published in German between 1807 and 1871 also returned some books in English. Likewise, a search for full-text books published in 1749 produced an initial result of 2,311 books, but after paging through the results, this ultimately resolved to 227 volumes, presumably through the progressive elimination of duplicate entries. Curiously, Latin was not available as a choice for limiting a search by language, though this will presumably be changed as increasing numbers of pre–1801 imprints are added to Google Books.

Result sets from Google Books must be examined with care. For example, a search for A. O. Abbott's *Prison Life in the South* with the result set restricted to full-view manifestations does not retrieve a matching manifestation in the first page of results but rather in the second. Google Books can likewise be unforgiving of faulty metadata. For example, a search for E. S. Abdy's *Journal of a Residence and Tour in the United States, 1833–34* produces a matching full-view manifestation only if one removes the "-34"—reduced from "1834" in the Boston Athenæum catalog—from the search argument: a case where "less is more" in a search argument.

Identifying a Manifestation in the Catalog of the Control Collections

Very few library catalogs have not undergone retrospective conversion to machine-readable form at some point over the last several decades. Such retrospective conversions often carry two very large caveats. They may represent the conversion of files other than the official catalog, and they may have been outsourced to private firms on terms that would convert the largest number of records at the lowest cost.

The records resulting from such conversion are often incomplete, sometimes to such an extent that the resource represented by the record cannot be identified with certainty. Perhaps the most famous (or notorious) example of this is the PREMARC records in the OPAC of the Library of Congress (LC), the products of a conversion of records from the LC's old (pre–AACR2) shelflist. PREMARC records have a high transcription error rate (which the LC estimates at 15 percent for call numbers), and the contractor instructions allowed for the routine omission of subtitles (interpreted broadly by the contractor), contents notes, and series, with results that were often less than helpful.[39] For example, browsing the LC online catalog under "Fourteenth Census of the United States" returns dozens of records with this title—and only this title—and on which the only distinguishing features are varying paginations and hints from the variety of subject headings assigned.

Similar conversions have taken place at most research libraries, including the ones that served as the control collection for this study. Fortunately, incomplete records in the online catalog of the control collections were only an occasional problem. Nevertheless, in cases where there was insufficient bibliographic data, the author needed to come up with a rule of thumb to determine whether a particular catalog record represented the manifestation being sought. Given that the object of the current study was to determine whether the number of matches in Google Books equaled or exceeded the number in the control collections, the benefit of the doubt was given to the control collections in these cases. The manifestation being described could not be determined in just three of these cases.

Finally, authority control remains imperfect among retrospectively converted records in online library catalogs, where records from different files—some authority controlled, some not—may have been merged. It was not uncommon when searching the catalog of a control collection to find the same person or corporate body represented by three or more headings, often differing from one another only very slightly. Unfortunately, while the differences were often barely noticeable to a human reader, a miss was as good as a mile to a machine, which duly segregated them under discrete headings.

Results

The author found digitized items (matching "full view" books) representing 235 entries in Google Books out of 398 in the sample from Cutter's catalog. The match rate for Google Books was 59.05 percent ± 4.83 percent. The corresponding match rate for the older research library (268 matches) was 67.34 percent ± 4.61 percent and for the later research library (162 matches) 40.7 percent ± 4.83 percent. All rates were significant at the 95 percent confidence level. The Z-tests comparing the Google Books results with those of each of the control libraries were significant in both cases (Google Books versus the older research library: Z = 2.351; confidence level > 95 percent; versus the later research library: Z = 5.106; confidence level > 99 percent). These data are shown in figure 1.

The sample included 98 entries representing pre–1801 imprints. Google Books and the two control collections differed markedly in their match rates for these materials. Only 27 were found in Google Books, while 66 were found in the "contemporary" research library and 48 were found in the "later" research library.

For post–1800 imprints, Google Books had a slightly higher match rate than the contemporary research library and a significantly higher rate than the later library (see figure 2). Out of 300 entries in the sample representing post–1800 imprints, 208 were found in Google Books while 202 were found in the contemporary library's OPAC (Z = -0.439; confidence level > 95 percent) and 114 in the later library's OPAC (Z = 7.613; confidence level > 99 percent). (See figure 2.)

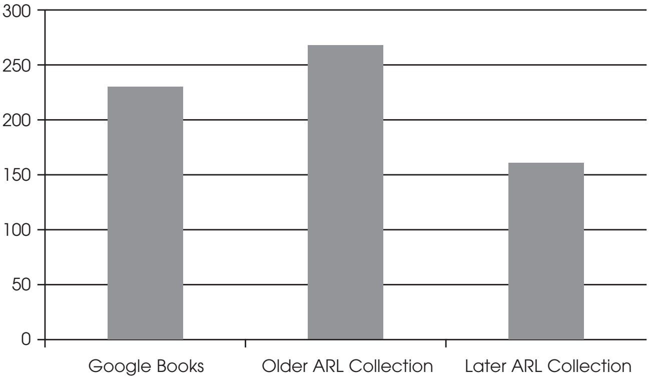

Figure 1. Number of Matches with Boston Athenæum Catalog Sample of 300 Titles

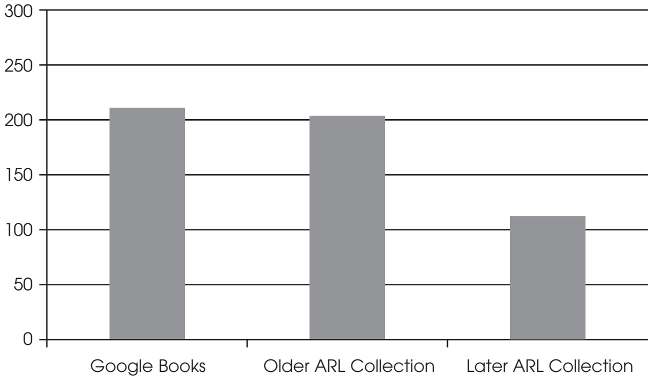

Figure 2. Number of Matches with 398 Post–1800 Sample Entries from Boston Athenæum Catalog

Discussion

The study employed a very strict definition of manifestation, requiring matching resources to conform in all bibliographic details with the entries in Cutter's Boston Athenæum catalog. This meant that in several cases a seemingly identical text—an expression in FRBR terms—was rejected as a match. Consequently, the match rates reported in this study should be treated as floors in terms of matching expressions.

In a similar vein, the author makes no judgments as to the continued value of the expression embodied in a given manifestation vis-à-vis a later expression, one that might have superseded that earlier expression in the Athenæum catalog. A library that owned a manifestation of that later expression but not the manifestation represented by the Athenæum catalog entry was deemed not to own a match. As an example of this phenomenon, the Athenæum catalog contained a four-volume collection of the works of William H. Seward published during his lifetime, while another library had a more complete set published posthumously. Presumably the later set was both more complete and more accurate from a scholarly perspective, but with no way to determine this objectively, it was rejected as a match in this study. Again, the result is an undercount of what scholars might consider to be matches.

Beyond these effects of the experimental design on match rates, some differences were the product of the design of the digitization projects themselves. The observed differences in Google Books match rates between pre–1801 imprints and post–1800 imprints is largely explained by the collections that are being selected for mass digitization. Initially, these have tended to be low-use items in remote storage facilities. American libraries often treat 1801 as a heuristic cutoff point for rare books, with books published before that year routinely segregated to rare book libraries or similar facilities not subject to mass digitization.[40] When pre–1801 imprints have been digitized, the volumes have occurred in the general collections, especially in the general collections of European partner libraries. Of the twenty-seven pre–1801 sample imprints found in Google Books, fourteen were digitized from European collections.

While the digitization of pre–1801 imprints would be attractive to Google in terms of increasing the comprehensiveness of Google Books, research libraries have less incentive. Many of the titles involved are already accessible online to these institutions on a subscription basis from vendors that have digitized preexisting proprietary microform collections of old books (Eighteenth Century Collections Online, Early American Imprints, etc.), so bringing them into the mass digitization stream would not necessarily increase the digital content available to their researchers. Of course, this does not rule out digitization of these materials at a later time.

One side effect of indiscriminately digitizing great swathes of the print collections of large research libraries is an increasing rate of duplication as more collections are digitized, and a decreasing marginal rate of return for each newly digitized collection. For example, Lavoie estimated that while 58 percent of the holdings (approximately 18 million) of the original Google 5 were unique to a single institution, this would likely be true for only 22 percent of the holdings (approximately 8 million) of an additional Google 5.[41]

However, increasing duplication can introduce potential benefits in some contexts. While the study encountered many volume-level duplications, the study itself was not designed to measure this duplication and made no attempt to do so. Under ordinary conditions, one would want to keep such duplication to a minimum. But in a world where the power and storage capabilities of computers continue to increase at a staggering rate, this approach may no longer be valid, and a certain amount of redundancy may actually present advantages. In a slightly different context—that of translating a text from one language to another—Garreau observed in *The Washington Post*,

> The explosion of the Web . . . has enabled a revolution. Like so many successful human approaches, it relies on brute force and ignorance. This method cares little for how any language works. It just looks—Rosetta stone fashion—at huge amounts of text translated into different languages by humans. (Dump decades of U.N. documents into the maw.) Then it lets the machine statistically express the probability that words in one language line up together in a fashion comparable to another set of words in another language.[42]

In the context of Google Books, item-level duplication feeds into this "brute force and ignorance" method of machine-based quality control, especially as an alternative to more labor-intensive page-by-page human quality control (now mainly the responsibility of the partner libraries). One can easily imagine, for example, a two-step probabilistic method whereby Google first identifies identical images (setting a very high bar for a match in terms of visual pattern recognition), then selects from them the image from which optical character recognition (OCR) has produced the most satisfactory rendering into plain text (taking this as evidence of a relatively clear image of the original). This would simultaneously both reduce the number of duplicate volumes in Google Books and the number of poorly scanned images that would otherwise require manual intervention. A simpler process might be used to enable the supplying of missing pages when comparing two otherwise identical sequences of page images.

From the point of view of scholarly research, poor image quality and occasional missing pages may be less of a problem pragmatically than it is absolutely (as a guarantor of textual integrity). Researchers—the intended audience of a research library—are typically interested in books not as artifacts of cultural heritage or even necessarily as integral texts, but rather as containers of certain desired content. Consequently, most of them read books—at least online—only to the extent necessary to extract the desired content, and this may not be reading as we know it. This is seen, for example, in a recent large-scale survey of British academics, where most respondents reported that they "dipped in and out of several chapters" when reading e-books rather than reading continuously.[43] Given this user behavior, Google's post–scanning priorities might not necessarily be in image quality in general (except to the extent that this affects the ability of their software to recognize and index text) but rather in the quality of those images that people actually view (or try to view), since this can affect advertising revenue. Again, one should note that post–scanning quality control is currently the responsibility of the partner libraries.

Conclusion

This study has shown that—with some caveats—the pre–1872 digitized content now available in Google Books approximates that content available via the online catalog of a generic major American research library, and indeed is probably superior for post–1800 imprints. Google Books has reached this point in a remarkably short time—less than five years after the announcement of its initial Library Project—and given the large number of research library partners that have since been recruited, it seems likely that Google Books will eventually (perhaps very soon) become the single largest source for this content.

On the negative side of the ledger, two significant caveats must be recalled. The digitized images of individual pages are not always reliable—poor scanning can occasionally be so extensive as to render a digitized volume unusable—and folded maps and other illustrative matter are routinely scanned in their folded state, rendering them useless for research. One can reasonably expect that these flaws will be corrected over time, at least for high-demand texts: The users of the texts will insist on it and, at any rate, the libraries involved are committed to it.[44] Measuring the extent of this problem was not within the scope of the current study, but an extremely useful future research project would try to do so.[45]

On the positive side, Google Books provides full-text indexing, something of incalculable value that would have been inconceivable had these volumes not been scanned. This indexing allows one to search both within individual volumes and across the entire collection, facilitating text-based research in general, but especially historical research and the comparison of variant texts. While this indexing is dependent in individual cases on the quality of the original page scan and the fidelity of the OCR rendering, in the aggregate the amount of hidden content that is thus exposed far exceeds the amount that remains hidden (or imperfectly rendered via OCR). Additionally, Google Books is serving as a huge laboratory for what is called "document image understanding"—the increasingly sophisticated probabilistic analysis of page images to facilitate indexing, interpretation, and other activities.[46]

As noted above, in the past, large collections of works in the public domain—especially older English language works—were microfilmed by commercial firms in collaboration with various research libraries. The resulting microform collections have subsequently been digitized, either by the firms that did the original microfilming or by successor firms, and made available on a subscription basis. As Google

Books and other mass digitization projects continue their progress through various research library collections, the viability of these preexisting collections may increasingly come into question as subscribing institutions weigh their annual use of these materials against the annual charges they pay for access.

Currently only a small fraction of the materials in Google Books—perhaps 15 percent—is thought to be in the public domain.[47] The great bulk is still protected by copyright, including a large but unknown number of so-called orphan works for which it is difficult or impossible to locate the current copyright holder.

These materials, many of which have been digitized in the course of the Library Project, were the object of class-action lawsuits brought against Google in 2005 by the Association of American Publishers and the Authors Guild.[48] The parties proposed a settlement of these lawsuits on October 28, 2008, but at the time of this writing the fairness of its terms is still to be determined by the U.S. District Court involved. On July 2, 2009, the U.S. Justice Department informed the court that it had opened an antitrust investigation into the settlement.[49]

Among librarians and researchers, the reaction to the proposed settlement was in some ways emblematic of the ambivalence felt by many who stand to benefit from Google's mass digitization. Harvard University and others have objected to the proposed settlement on the grounds that it would grant a de facto monopoly to Google. Robert Darnton, director of the Harvard University Library, summarized his misgivings in the *New York Review of Books*:

> Google is not a guild, and it did not set out to create a monopoly. On the contrary, it has pursued a laudable goal: promoting access to information. But the class action character of the settlement makes Google invulnerable to competition. Most book authors and publishers who own US copyrights are automatically covered by the settlement. They can opt out of it; but whatever they do, no new digitizing enterprise can get off the ground without winning their assent one by one, a practical impossibility, or without becoming mired down in another class action suit. If approved by the court—a process that could take as much as two years—the settlement will give Google control over the digitizing of virtually all books covered by copyright in the United States.[50]

Given the research benefit that would accrue from providing direct integrated access to copyrighted material via Google, some mutually acceptable arrangement is likely to be reached, though its ultimate shape is hard to fathom at this point. The ramifications of any settlement are such that a lengthy court review seems likely.

Beyond the terms of the proposed settlement lies the larger question of how Google Books will ultimately affect the world of learning. By making so much of the printed record available in digital form—and so rapidly—Google Books is both transforming how scholars use the printed texts of the past and feeding a larger fundamental reshaping of the world of scholarly research. Fortunately for the author, speculation on the ramifications of these changes lies beyond the scope of the current study. But Google Books clearly is already having a dramatic effect, both on libraries and on scholarship. Indeed, a significant number of the sources cited in this study—beyond the objects of the study itself—were consulted in Google Books rather than in a physical library. As more and more scholarly research is conducted online first—and especially if the universe of digitized copyrighted works is ultimately opened up in Google Books—libraries may find that researchers are not linking out of online catalogs to versions of works available on Google Books but are rather linking to library catalogs for those cases where a version available on Google Books is not satisfactory for their purposes. We will then be entering a brave new world for both research and libraries.

References and Notes

1. Google, "Google Checks Out Library Books," press release, Dec. 14, 2004, www.google.com/press/pressrel/print_library.html (accessed Sept. 18, 2009).
2. For an example, see the agreement between the University of Michigan and Google, "Cooperative Agreement" (n.d.), www.lib.umich.edu/mdp/um-google-cooperative-agreement.pdf (accessed Sept. 18, 2009).
3. Karen Coyle, "Google Books as Library," online posting, Nov. 22, 2008, Coyle's InFormation, http://kcoyle.blogspot.com/2008_11_01_archive.html (accessed Sept. 18, 2009).
4. See, for example, Committee on Institutional Cooperation and Google, "Cooperative Agreement" (n.d.), www.cic.net/Home/Projects/Library/BookSearch/CIC-Google.aspx (accessed Sept. 18, 2009), which commits to digitizing "not less than 10,000,000 volumes" (2); University of California and Google, "Cooperative Agreement" (n.d.), www.cdlib.org/news/ucgoogle_cooperative_agreement.pdf (accessed Sept. 18, 2009), which commits to digitizing "no less than two and a half million (2,500,000) volumes" (2); University of Texas at Austin and Google, "Cooperative Agreement" (n.d.), www.lib.utexas.edu/sites/default/files/google/utexas_google_agreement.pdf (accessed Sept. 18, 2009); and University of Virginia and Google, "Cooperative Agreement" (n.d.), www2.lib.virginia.edu/press/uvagoogle/pdf/Google_UVA.pdf (accessed Sept. 18, 2009). The agreements typically include previously digitized content, according to Europeana, About Us: Background, www.europeana.eu/portal/aboutus.html#background (accessed Sept. 18, 2009).
5. Miguel Helft, "Microsoft Will Shut Down Book Search Program," *New York Times*, Technology section, May 24,

2008, www.nytimes.com/2008/05/24/technology/24soft.html (accessed Sept. 18, 2009).

6. Google, “Google Checks Out Library Books,” press release, Dec. 14, 2004, www.google.com/press/pressrel/print_library.html (accessed Sept. 18, 2009); Google, Corporate Information: Company Overview, www.google.com/intl/en/corporate (accessed Sept. 18, 2009).
7. Michael A. Keller, University Librarian, Stanford University, quoted on Google Book Search Library Partners, http://books.google.com/googlebooks/partners.html (accessed Sept. 18, 2009).
8. The University of Michigan’s workflow is illustrated in a flow chart, “Michigan Digitization Project Workflow,” www.lib.umich.edu/files/services/mdp/MDPflowchart_v3.pdf (accessed Sept. 18, 2009).
9. International Federation of Library Associations and Institutions, Section on Acquisition and Collection Development, “Guidelines for a Collection Development Policy Using the Conspectus Model,” (2001), http://archive.ifla.org/VII/s14/nd1/gcdp-e.pdf (accessed Sept. 18, 2009).
10. The items are IC-3, “The printed library materials I need for my work,” IC-4, “The electronic information resources I need,” and IC-8, “The print and/or electronic journal collections I require for my work.” On the 2008 survey, the gap on IC-3 slipped back into the positive territory it had occupied in 2002, but the importance of this is difficult to determine. Standard deviations on these items are large—typically between 1.9 and 2.2—making detecting and measuring trends difficult. The ARL LibQual+ notebooks for 2002 to the present can be examined at www.libqual.org/Publications/all.cfm?PubType=11 (accessed Sept. 18, 2009).
11. Walt Crawford, “Libraries and Google/Google Book Search: No Competition!” online posting, June 21, 2006, Google Librarian Central, www.google.com/librariancenter/articles/0606_03.html (accessed Sept. 18, 2009). Thomas E. Nisonger’s *Collection Evaluation in Academic Libraries* (Englewood, Colo.: Libraries Unlimited, 1992) contains annotated bibliographies on overlap in general and in the context of the Research Libraries Group Conspectus.
12. Brian Lavoie, Lynn Silipigni Connaway, and Lorcan Dempsey, “Anatomy of Aggregate Collections: The Example of Google Print for Libraries,” *D-Lib Magazine* 11, no. 9 (Sept. 2005), doi:10.1045/september2005-lavoie, www.dlib.org/dlib/september05/lavoie/09lavoie.html (accessed Sept. 18, 2009).
13. Brian F. Lavoie and Roger C. Schonfeld, “Books without Boundaries: A Brief Tour of the System-wide Print Book Collection,” *Ubiquity* 7, no. 37 (Sept./Oct. 2006), doi:10.1145/1167867.1167868, www.acm.org/ubiquity/views/v7i37_books.html (accessed Sept. 18, 2009).
14. University of Michigan and Google, “Cooperative Agreement.”
15. Coyle, “Google Books as Libraries.”
16. Lavoie, Connaway, and Dempsey, “Anatomy of Aggregate Collections.”
17. British Library, Help for Library Researchers, English Short Title Catalogue—Introduction, www.bl.uk/reshelp/findhelprestype/catblhold/estcintro/estcintro.html (accessed Sept. 18, 2009).
18. Eltjo Buringh and Jan Luiten van Zanden, “Charting the ‘Rise of the West’: Manuscripts and Printed Books in Europe, a Long-Term Perspective from the Sixth through Eighteenth Centuries,” *The Journal of Economic History* 69, no. 2 (June 2009): 409–45, doi:10.1017/S0022050709000837, http://dx.doi.org/10.1017/S0022050709000837 (accessed Sept. 18, 2009).
19. Lavoie and Schonfeld, “Books without Boundaries.”
20. Robert B. Townsend, “Google Books: What’s Not to Like?” online posting, April 30, 2007, AHA Today, http://blog.historians.org/articles/204/google-books-whats-not-to-like (accessed Sept. 18, 2009); Ronald G. Musto, “Google Books Mutilates the Printed Past,” *The Chronicle of Higher Education* 55 (June 12, 2009): B4, http://chronicle.com/article/Google-Books-Mutilates-the/44463 (accessed Sept. 18, 2009).
21. Karen Coyle, “Mass Digitization of Books,” *Journal of Academic Librarianship* 32, no. 6 (Nov. 2006): 641–45, http://dx.doi.org/10.1016/j.acalib.2006.08.002 (accessed Feb. 9, 2010).
22. *Public Libraries in the United States of America* (Washington: GPO, 1876): 1:760, www.archive.org/details/publiclibrariesi00unitrich (accessed Sept. 18, 2009). The term *public libraries* is used very broadly in the nineteenth century, including both academic and private research libraries. Curiously, part 1 of this work is not present in Google Books aside from the chapter on theological libraries digitized from the Andover-Harvard Theological Library. A copy identifying itself as part 1 of a University of Michigan copy in Google Books is, in fact, a digitization of part 2.
23. Charles C. Jewett, *Notices of Public Libraries in the United States of America* (Washington, D.C.: Smithsonian Institution, 1851): 192, http://books.google.com/books?id=paMnAAAAMAAJ (accessed Sept. 18, 2009).
24. Jewett, *Notices of Public Libraries,* 762–73.
25. *Catalogue of the Library of the Boston Athenæum, 1807–1871* (Boston: The Athenæum, 1874–1882).
26. Prospectus quoted in Josiah Quincy, *The History of the Boston Athenæum* (Cambridge, Mass.: Metcalf and Co., 1851): 12, http://books.google.com/books?id=RHHRoEiGRdgC (accessed Sept. 18, 2009).
27. Jewett, *Notices of Public Libraries*, 22.
28. Edward Edwards, *Memoirs of Libraries* (London: Trübner & Co., 1859): 2:194, http://books.google.com/books?id=TH4NAAAAQAAJ (accessed Sept. 18, 2009).
29. James Lyman Whitney, “Considerations as to a Printed Catalog in Book Form for the Boston Public Library,” *The Library Journal* 24 (July 1899): 10, http://books.google.com/books?id=ggcCAAAAYAAJ (accessed Sept. 18, 2009).
30. *Catalogue of the Library of the Boston Athenæum, 1807–1871*.
31. Anthony Panizzi, “Rules for the Compilation of the Catalogue,” *Catalogue of Printed Books in the British Museum*, (London, 1841): 1:[v]–ix; Charles Ammi Cutter, “Explanations,” *Catalogue of the Library of the Boston Athenæum, 1807–1871* (Boston: The Athenæum, 1874–1882): 1:page following title page, http://books.google.com/books?id=Zbjcm5OGlsIC (accessed Sept. 18, 2009).
32. IFLA Study Group on the Functional Requirements for

Bibliographic Records, *Functional Requirements for Bibliographic Records: Final Report*, UBCIM Publications-New Series, vol. 19 (Munich: K.G. Saur, 1998): 17–25, www.ifla.org/files/cataloguing/frbr/frbr.pdf (accessed Mar. 2, 2010).

33. Ibid., 74.
34. Assocation of Research Libraries, ARL Library Data Tables 2007–08, worksheet "coll1," www.arl.org/bm~doc/08tables.xls (accessed Sept. 18, 2009).
35. IFLA Study Group, *Functional Requirements for Bibliographic Records*, 74.
36. Charles Evans, *American Bibliography: A Chronological Dictionary of All Books, Pamphlets, and Periodical Publications Printed in the United States Of America from the Genesis of Printing in 1639 Down to and Including the Year 1820* (New York: P. Smith, 1941–67); Early American Imprints, Series I: Evans, 1639–1800, www.readex.com/readex/product.cfm?product=247 (accessed Sept. 18, 2009).
37. A Z-test is a statistical test used to calculate the probability that a given sample result lies within the range of results that one would expect if the only effect present was the operation of chance. A results outside this range suggests that something other than the operation of chance is contributing to that result.
38. UNESCO, Revised Recommendation Concerning the International Standardization of Statistics on the Production and Distribution of Books, Newspapers and Periodicals (Nov. 1, 1985), http://portal.unesco.org/en/ev.php-URL_ID=13146&URL_DO=DO_TOPIC&URL_SECTION=201.html (accessed Sept. 18, 2009).
39. Thomas Mann, *The Oxford Guide to Library Research* (New York: Oxford Univ. Pr., 1998): 185–87.
40. For example, the original agreement with the University of Michigan explicitly excluded special collections materials. "Cooperative Agreement" (n.d.) 1.1, www.lib.umich.edu/mdp/um-google-cooperative-agreement.pdf (accessed Sept. 18, 2009).
41. Brian Lavoie, "Anatomy of Aggregate Collections: Exploring Mass Digitization and the 'Collective Collection,'" PowerPoint presentation, NELINET, Sept. 21, 2006, www.oclc.org/research/presentations/lavoie/nelinet2006.ppt (accessed Sept. 18, 2009).
42. Joel Garreau, "Tongue in Check: With Translation Technology on Their Side, Humans Can Finally Lick the Language Barrier," *Washington Post*, Style section, May 24, 2009, www.washingtonpost.com/wp-dyn/content/article/2009/05/21/AR2009052104697.html (accessed Sept. 18, 2009).
43. David Nicholas et al., "UK Scholarly E-Book Usage: A Landmark Survey," *Aslib Proceedings* 60, no. 4 (2008): 311–34.
44. See, for example, the HathiTrust instructions at http://babel.hathitrust.org/cgi/mb?a=page;page=help (accessed Sept. 18, 2009).
45. Reports of this problem to date have been anecdotal. See, for example, Robert B. Townsend, "Google Books: Is It Good for History?" *Perspectives Online* 45, no. 6 (Sept. 2007), www.historians.org/perspectives/issues/2007/0709/0709vie1.cfm (accessed Sept. 18, 2009); Musto, "Google Books Mutilates the Printed Past."
46. See, for example, Faisal Shafait et al., "Background Variability Modeling for Statistical Layout Analysis," *Proceedings of the 19th International Conference on Pattern Recognition, December 8–11, 2008, Tampa, Florida, USA*, (2008) IEEE Computer Society, 2008, doi:10.1109/ICPR.2008.4760964, http://dx.doi.org/10.1109/ICPR.2008.4760964 (accessed Sept. 18, 2009).
47. Geoffrey Nunberg, "Google's Book Search: A Disaster for Scholars," *Chronicle of Higher Education* 56, no. 1 (Aug. 31, 2009): B4, http://chronicle.com/article/Googles-Book-Search-A/48245 (accessed Sept. 18, 2009).
48. Jonathan Band, "A Guide for the Perplexed: Libraries and the Google Library Project Settlement" (Nov. 13, 2008), http://wo.ala.org/gbs/wp-content/uploads/2008/12/a-guide-for-the-perplexed.pdf (accessed Sept. 18, 2009); Jonathan Band, "A Guide for the Perplexed Part II: The Amended Google–Michigan Agreement" (June 12, 2009), http://wo.ala.org/gbs/wp-content/uploads/2009/06/google-michigan-amended.pdf (accessed Sept. 18, 2009).
49. William F. Cavanaugh, Deputy Assistant Attorney General, letter to Hon. Denny Chin, July 2, 2009, http://graphics8.nytimes.com/packages/pdf/technology/20090702_GOOGLE_DOJLetter.pdf (accessed Sept. 18, 2009).
50. Robert Darnton, "Google and the Future of Books," *The New York Review of Books* 56, no. 2 (Feb. 12, 2009), www.nybooks.com/articles/22281 (accessed Sept. 18, 2009). A subsequent response to Darnton is Paul N. Courant et al., "Google and Books: An Exchange," *The New York Review of Books* 56, no. 5 (Mar. 26, 2009), www.nybooks.com/articles/22496 (accessed Sept. 18, 2009).