WHAT’S MISSING FROM GOOGLE BOOKS?

Michele Seikel¹ and Thomas Steele²

¹Associate Professor/Cataloger, Edmon Low Library, Oklahoma State University, Stillwater, Okla. 74078
²Assistant Professor/Science and Technology Cataloger, Bizzell Memorial Library, University of Oklahoma, Norman, Okla. 73019

ABSTRACT
Points out that maps, plates, and plans have been a staple within published materials in the fields of history, geology, geography, and land use (agriculture), all of which rely heavily on maps, for two hundred years. A study published in 2010, entitled “Google Books as a General Research Collection,” identified a problem encountered with digitized versions of monographs that include folded illustrations. The article pointed out that many titles in Google Books could have been digitized without unfolding their accompanying illustrations, with the help of the industrial scanning process, used to digitize huge numbers of titles quickly. In order to determine the extent of the problem, this study attempts to one, compare the numbers of missing or only partly-digitized folded illustrations within titles that are fully available for viewing in Google Books and two, other large databases of digitized monographs, Open Library and Project Gutenberg. Indicates that Google Books generally scanned oversized or detached folded illustrations, or else omitted them, whereas Open Library’s scanning process, more often, correctly scanned folded materials.

KEYWORDS: Google Books, Project Gutenberg, Open Library, Internet Archive

Introduction
Although a flurry articles and blog entries have been published since 2007 concerning Google Books, many of them were opinion pieces, expressing the fear that the enormous size of the Google Books database would discourage other digitization projects or simply replace libraries altogether. Several journal articles focused on the effect of the mass digitization project on copyright issues in the United States and other western countries. However, a survey of the database published in Library Resources & Technical Services stated that “… 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” (Jones, 2010, p. 79). In addition, Robert Townsend, writing in his American History Association Today blog (Townsend, 2007), gave a detailed analysis of problems with poor scan quality, faulty metadata and lack of access to some materials already in the public domain.

In an issue of First Monday, Paul Duguid (2007) compared the results of a title search in Project Gutenberg with the same search results in Google Books. By describing this simple process in considerable detail, Duguid was able to give a vivid picture of the problems related to scanning. As he stated, “The Greek text, the footnotes, the black
The digitization Problem

Google spent the early 2000’s assembling a monumental database of digitized monographs chosen from the collections of several largest research libraries in the western world. As the world now knows, the result was Google Books, which include millions of titles in several languages. In order to accomplish this tremendous project, Google used an industrial scanning strategy, which allowed each title to be scanned very rapidly. “In its frenzy to digitize the holdings of its partner collections, in this case those of the Stanford University Libraries, Google Books has pursued a “good enough” scanning strategy. The pages of the books were hurriedly reproduced; without employing any quality control, either during or after scanning” (Jones, 2010, p. 79). In many cases, “good enough” produced a readable, conveniently online product. At times, however, readers have reported missing or poorly scanned pages, particularly if they are of different sizes or if they are separate from the main text. “Similarly, the problem of non-textual content (illustrations, maps, etc.) in the digitized volumes, while not significant for indexing, is significant for research, especially in the case of folded materials, which Google scanners typically digitize in its folded form” (Musto, 2009, p. B4).

Apparently, the original Digitization Project Plan did not account for differently sized pages or for the pages separate from the main text (in a pocket, for instance). In a normal digitization plan, a method of accounting for such materials would be part of the workflow. “… libraries will find that some items are either too fragile to be put through the mass digitization process or are too far from the norm to be suitable to that process. Some books will be too large or too small; others will have odd-sized plates or folded maps that will need special handling. So digitizing an entire library will require some mass digitization and some special digitization projects” (Coyle, 2006).

Roughly within the same time period as the development of Google Books, the Internet Archive and Project Gutenberg have developed smaller digital book databases, although both of these are “open” to donations from the public, and have even had some donations of texts originally digitized by Google Books. While there have been some complains about Project Gutenberg’s early use of ASCII text, however, this smaller project did never garner the publicity or notoriety that the breadth of the Google projects did. Writing in 2007, Paul Duguid compared the quality of Project Gutenberg’s digitization with that of Google Books, using a famously difficult text, *Tristram Shandy*, as the point of comparison. He concludes, “Thus, where scanning should conquer the disparaged ASCII transcription of Project Gutenberg, in this case, the text chosen by the scanning project and presented as the first choice of the reader is so inadequate that there is little improvement for the ordinary reader, who at least can read the ASCII text (Duguid, 2007).”

The subject matter of titles, which includes folded illustrations, varies much more widely than expected; including geography, travel, history, ethnology, climatology, mineralogy, archeology, anthropology, urban planning, hydrology, geology, agriculture and sea
navigation. Scholars, working in the fields where maps and other folded illustrations may contain vital data have commented on their absence or poorly scanned state in Google Books. “And like a number of the other books on the site, some pages appear to have been scanned in mid-transit through the scanner, while pages and tables too large to fit within their imaging constraints were simply cut off. This makes much of the text unreadable and presumably less discoverable” (Townsend, 2007).

Methodology
An initial sample of 306 monograph titles was drawn from Index to Maps in Books and Periodicals (American Geographical Society of New York. Map Dept., 1968), a ten-volume catalogue which includes titles in several languages and describes maps from around the world. The Index specifies both the publication date and the titles in which the maps are included. Titles that had publication dates later than 1929 were not included in project. This is because it is much more likely that materials are already in the public domain (i.e., published no later than the 1920’s) will be available in full-text; 1923 is the year beyond which materials are clearly in the public domain, according to American copyright law. “Using the year 1923 as a rough break-off point between materials that are out of copyright and materials that are in copyright … more than 80 percent of the materials in the Google 5 collections are still in the copyright” (Lavoie, Connaway, Dempsey, 2005). This limited the titles that we could choose to search for in Google Books, but improved the odds that they would actually have full-text available. Our search of the ten-volume Index was not exhaustive, but included random pages from all of the volumes. An additional list of 55 titles was compiled from random pages of the 4-volume Research Catalog of Maps of America to 1860 in the William L. Clements Library (Marshall, 1972).

Each of the title on both lists was then searched in OCLC WorldCat, in order to verify the existence of folded illustrations via the physical description portion of the catalogue records. Below are some examples of such typical physical descriptions in the catalogue records:

```
xxvii, 498 p. front., illus., plates, maps (part fold.) 27 cm.
```

```
Nine folded maps in pocket.
```

Only if such verification was found in OCLC, that a title was added to the lists and searched in Google Books. To provide a comparison sample, the same titles were then searched in the same way, in two other databases- Open Library, which is the creation of the Open Content Alliance and the Internet Archive, and Project Gutenberg. Open Library currently has about 1.6 million full-text books within a much larger online bibliography of 23 million titles. Project Gutenberg currently has about 33,000 downloadable books in its own database and another 100,000 digitized titles in affiliated databases.

Results of the search
Of the 359 monograph titles on our lists, a total of 151 were found to be available for viewing in their entirety in Google Books. The other 208 monograph titles were only available for previews in snippets, not available for preview, or not found in the database. From the selection of 151 monograph titles that were fully available for viewing, only two had every map and folded illustration scanned.
Table 1: Google Books

<table>
<thead>
<tr>
<th>Total Titles Searched in Google Books</th>
<th>Snippets or Not Available to View</th>
<th>Missing or Partly Scanned Illustrations</th>
<th>Correctly Scanned on Adjacent Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>359</td>
<td>208</td>
<td>149</td>
<td>2</td>
</tr>
</tbody>
</table>

However, those two titles appeared not to have any folded maps, but actually maps that covered adjacent pages, with the spine of the book dividing the maps. It is questionable if those maps should have been described as folded in the physical description portions of their catalog records. For the remaining 149 titles, the maps and illustrations were either scanned folded, as shown in the illustration below, or missing from the book entirely. Maps described in the catalog records as being in pockets usually were missing.

Unlike Google Books, Open Content Alliance invites the public to contribute entries and/or digitized titles to its database. Therefore, some of the entries found in Open Library were originally digitized by Google Books. Others were digitized from microfiche copies. Internet Archive itself has also digitized many of the titles that are viewable in the Open Library. Its own digitization process usually produces correctly scanned maps, plans, and other illustrations, including those which are oversized, such as the title below.
In Open Library, the quality of scanning varies considerably, depending on the origin of the scanned copy. Those titles which have been digitized in the Internet Archive’s own scanning operation are correctly scanned, including maps which are oversized. Titles that were digitized from a microform copy routinely have missing illustrations. Titles that were originally digitized by Google and then donated to Open Library had generally missing illustrations, including maps. Frequently, the captions for illustrations were on the pages, but the illustrations themselves were missing.

<table>
<thead>
<tr>
<th>Total Titles Searched in Open Library</th>
<th>Not Available to View</th>
<th>Missing or Partly Scanned Illustrations</th>
<th>Correctly Scanned Illustrations</th>
</tr>
</thead>
<tbody>
<tr>
<td>359</td>
<td>172</td>
<td>128</td>
<td>59</td>
</tr>
</tbody>
</table>

For Project Gutenberg, while the text itself was in ASCII, two of the 359 titles searched did have complete, correctly scanned images. Unfortunately, only a total of four titles from our lists were found in the Project Gutenberg database, making the results inconclusive for lack of a significant sample size.

Conclusion
The difficulty of finding titles with folded illustrations within fully-viewable texts meant that more than half of the works from our compiled list could not have been found. But we succeeded in ascertaining that, of the titles with folded illustrations found in the Google Books database, all were either scanned in their folded state or missing entirely. Our
sample size is large enough to indicate that illustrations which were either differently sized from the text or separated in a pocket were not included in the digitization workflow as a general rule. They remained unscanned or were scanned folded, and are not available to the Google Books readers now.

Open Library, the only other book digitization database we examined with a sample size as significant as Google Books, had better quality and more complete scanning of those titles originally digitized by Internet Archive. Project Gutenberg’s results were inconclusive, even though the fact that it did have complete images scanned for at least two titles leaves open the possibility of better image reproduction than Google Books.

The reason, why this matter remains worth examining, is that the sheer size and centrality of Google Books causes people to search there first, rather than look in other, smaller book databases. “With each scanned page, Google Books’ Library Project, by its quantity if not necessarily by its quality, makes the possibility of a better alternative unlikely” (Duguid, 2007). One can imagine a scenario for an individual work in which the Google Books copy might become the only copy available to researchers of the future, who can no longer obtain access to a printed copy. In such a scenario, the unscanned or mis-scanned map, table, or chart would be lost forever, regardless of its value to researchers.

REFERENCES


