An Assessment of Image Quality in Geology Works from the *HathiTrust* Digital Library

Scott R. McEathron
T. R. Smith Map Collection
University of Kansas Libraries
1301 Hoch Auditoria Dr.
Lawrence, KS 66045-7537
macmap68@ku.edu

Abstract

This study assesses the quality of both images and text in a sample from the 2,180 works on geology from the *HathiTrust* Digital Library (multi-institutional digital repository)--an outgrowth of the Michigan Digitization Project and partnership with Google, Inc. A random sample of 180 (consisting of 47,287 pages) was made and reveals many patterns and characteristics of the digital manifestations of these works. The good news is that of the total 47,287 pages that were reviewed, only 2.5% had errors. The bad news, of the 180 works, 114 or 63% had at least one scanning error. It is important for librarians and readers to know the strengths and shortcomings of this repository in considering future decisions on both de-accessioning and remote storage of works from libraries.

Introduction

Partnering with libraries and publishers, Google, Inc. has created the World's largest digital collection and index of books and journals. The broad implications of how this digital collection may transform future access and use of the works it contains, and the subsequent future of libraries has been the focus of several articles and opinion pieces of late (Dougherty, 2010; Jones, 2010; Nunberg, 2009; Darton, 2009). However, much of what has been written has also focused on the Google Book settlement with the Authors Guild, and the Association of American Publishers (Proskine, 2006; Band, 2009; Okerson, 2009). A few articles have begun making assessments of image quality and the means of access used within the Google Book product (James, 2010; Duguid, 2007; Townsend, 2007). However, these articles have been very limited in scope or in the size of their samples. Studies by Duguid (2007) and Townsend (2007) have limited their assessments to a single work. James's found less than 1% of the pages in his

sample had a significant error. However, the study had a relatively small sample of only 2,500 pages from 50 works.

The aim of this study is to assess the quality of both images and text in a sample from the 2,180 works on geology from the *HathiTrust* Digital Library (multi-institutional digital repository)--an outgrowth of the Michigan Digitization Project and partnership with Google, Inc. (HathiTrust, n.d.). The *HathiTrust* has become a primary repository for much of the digitization happening at Committee on Institutional Cooperation (CIC) and the University of California system libraries for the Google Book project. While this study specifically makes an assessment of the *HathiTrust* Digital Library, since much of the content is the same, many of the conclusions may, by extension, also be valid for portions of the Google Book project.

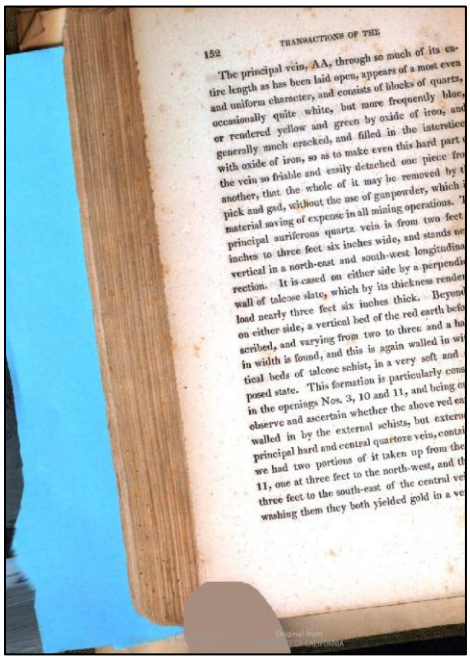
Methodology

All records for works that are fully available within the *HathiTrust* Digital Library and indexed with the subject term "geology" were downloaded into *Endnote* from the University of Michigan Libraries' online catalog: "Mirlyn." A total of 2,180 works met these criteria as of March 12, 2010. A random sample of 180 works was made from the total population of 2,180. Data gathered from the sample included: title; author; format; number of standard illustrations within the work and the number of standard illustrations with scanning errors; the number of large format illustrations (foldouts) and number of large formats with scanning errors; number of pages of each work and the number of pages of text with scanning errors; date published and original owning library of source document. A standard illustration was considered any image (i.e. woodcut, lithograph, and photograph) that was not a foldout or oversized illustration and kept in a back pocket. A scanning error was simply whether or not the illustration was capable of communicating the information it was intended to. Missing images were also considered an

error. Most illustrations are degraded to some extent in the digitization process. For the purposes of this study, they were judged using a pass or fail criteria: either they were adequate or they were not. Similar criteria for pages of text were also used: if the page was missing, unreadable or missing words or information that made it unreadable, it was considered a scanning error.

Results

A total of 47,287 pages of text were evaluated. Of that, 865 pages, or 1.8% were missing or deemed to be scanning errors in the text. Of the 180 works in the sample, 34 works or 19% had at least one scanning error of text. One work, *An elementary treatise on mineralogy and geology, designed for the use of pupils*, originally published in 1822, was missing 566 pages. This accounted for 65% of all the text scanning errors. The remaining errors were mostly poor scans of pages (Figure 1).


152 TRANSACTIONS OF THE

The principal vein, AA, through so much of its entire length as has been laid open, appears of a most even and uniform character, and consists of blocks of quartz, occasionally quite white, but more frequently blue, or rendered yellow and green by oxide of iron, and generally much cracked, and filled in the interstices with oxide of iron, so as to make even this hard part of the vein so friable and easily detached one piece from another, that the whole of it may be removed by the pick and gad, without the use of gunpowder, which is a material saving of expense in all mining operations. The principal auriferous quartz vein is from two feet six inches to three feet six inches wide, and stands nearly vertical in a north-east and south-west longitudinal direction. It is cased on either side by a perpendicular wall of talcose slate, which by its thickness renders the load nearly three feet six inches thick. Beyond this on either side, a vertical bed of the red earth before described, and varying from two to three and a half in width is found, and this is again walled in with vertical beds of talcose schist, in a very soft and decomposed state. This formation is particularly conspicuous in the openings Nos. 3, 10 and 11, and being enabled to observe and ascertain whether the above red earth was walled in by the external schists, but external to the principal hard and central quartzose vein, containing gold, we had two portions of it taken up from the openings 11, one at three feet to the north-west, and the other three feet to the south-east of the central vein, and washing them they both yielded gold in a very

Figure 1. – Example of text scanning error.

A total of 8,098 standard images were contained with the 180 sample works. Of the total, 98 or 1.2% were missing or deemed to be scanning errors. Of the 180 works in the sample, 35 or 19% had at least one scanning error of a standard illustration. The work with the most errors, numbering thirteen, was *Ground-water hydrology, historical water use, and simulated ground-water flow in Cretaceous-age Coastal Plain aquifers near Charleston and Florence, South Carolina* (1996). One problem identified that caused errors in both text and images may be a result of the automatic quality control processing of the page images--resulting in images or parts of text being clipped out (Figure 2). However, this did not seem to be a problem in the Google book interface for these same works (as the images had already been reprocessed to correct these errors).

Gneiss-Studie. 13

müsste aber jedenfalls vor Ablagerung der ebenfalls palaeozoischen Kalkschiefer und carbonischen Grauwacken eingetreten sein, da diese abwechselnd theils auf Glimmerschiefer, theils auf dem Gneiss liegen, ohne selbst die geringsten Spuren einer Contactmetamorphose zu zeigen.

Der Gneiss wäre danach also altpalaeozoischer Granit, der sich von dem echten mesozoischen Granit sowohl durch seine Structur als auch durch die ungeheure Mächtigkeit seines Contacthofes erheblich unterschiede. Ich habe in diesem Gebiete nicht eine einzige Thatsache auffinden können, welche einer solchen Auffassung das Wort reden könnte, wohl aber viele, die dagegen sprechen. Ich will die vier hauptsächlichsten anführen.

1. Nirgends entsendet, wie schon erwähnt, der Gneiss Apophysen in sein Dach, eben so wenig treten im Glimmerschiefer Gneissgänge auf. Es fehlt also dieser wichtige Beweis für das jüngere Alter und die eruptive Natur des Gneisses ganz.

2. Die Grenze des Gneisses gegen den Glimmerschiefer fällt stets genau in die Richtung der Schichtung in letzterem, so dass die jüngeren oberen Lagen des Glimmerschiefers oder die Einlagerungen von Amphiboliten und Dolomiten nirgends direkt auf den Gneiss zu liegen kommen. Ebenso ist die Schichtung oder Bankung im Gneiss zu derjenigen im Glimmerschiefer parallel, so dass sich beide wie ältere und jüngere Theile eines concordanten Schichtensystemes verhalten und der Gneiss durchaus als der Untergrund erscheint, auf dem die Glimmerschiefer sich abgelagert haben.

Fig. 1. Gneissgang im Gneiss der Teufelsklipp. (½ nat. Gr.)

3. Im Gneiss ist eine so stark ausgeprägte kataklastische Structur durchaus nicht zu beobachten, wie sie doch nothwendig entstanden sein müsste, wenn er durch Dynamometamorphose aus Granit hervorgegangen wäre.

4. Klastische Bestandtheile, wie sie in den sicher palaeozoischen Ablagerungen stets vorkommen, sind in den Glimmerschiefern nicht nachweisbar und waren deshalb wohl auch nie darin vorhanden.

Es fehlt also jeder Grund, um in ihnen einen palaeozoischen Contacthof zu sehen. Damit soll selbstverständlich keineswegs das Vorkommen lokaler Contacthöfe im Glimmerschiefer geleugnet werden, wie

Figure 2. – Example of standard illustration scanning error (Figure missing).

A total of 223 foldouts or large format illustrations were contained within the physical works of the sample of 180. Since all were missing from the digital version--all were counted as scanning errors. Obviously, there was a conscious decision by Google not to digitize foldout and large format illustrations (no doubt to increase the speed of scanning). Of the 180 works in the sample, 77 or 42% had at least one foldout or large format illustration. Thus for geology works, we can infer that by not scanning the foldouts or large format illustrations results in scanning errors 42% of the time.

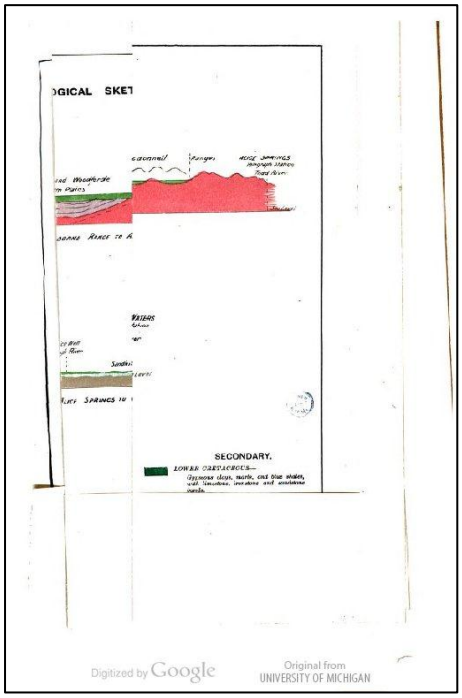


Figure 3. – Example of foldout scanning error.

Discussion

When the different types of errors are taken all together, within the sample of 180 works, there were a total of 1,186 scanning errors. Thus, of the 180 works, 114 or 63% had at least one scanning error. Google has classed errors into two forms: material and processing (York, 2010). Material errors are the result of deficiencies in the physical works (i.e. missing pages). Processing errors are those which result from the post-scan processing of the image. Of course

there are also the human errors associated with the procedure of manually turning the page (hand in the picture). This study suggests a fourth type of error; "policy" error. In order to achieve the massive scale deemed necessary for the project to be successful, the scanning of larger format foldouts where originally neglected. This policy can result in a large number of errors; especially in the case in works from certain disciplines such as geology, since a large percentage of works contain large foldout illustrations.

The policy of not scanning large format foldouts has implications of quality and completeness for the *HathiTrust* Digital Library and the Google Book product. The foldouts are of central importance for many works. Why would the original publisher go through the expense of compilation, printing them if they were not? In fact, for many works, they are the central intellectual work--the text is ancillary to the map. For example, Robert Bailey's *Ecoregions of the United States*, the original map was published in 1976 and the explanatory text to the map *Description of the Ecoregions of the United States* was not published until 1978 (Bailey, 1978). The central element of the work in this example is the map.

It should be pointed out that the post processing of the images has continued to improve. Thus, when the images are reprocessed, many of the errors are corrected (York, 2010). Thus, the results of this paper are really just a "snapshot in time" of how the images appeared in the summer of 2010 when this research was conducted. Also, the policy of not scanning large foldouts may change if it already has not. This will eventually result in fewer percentages of scanning errors within the texts and illustrations. Given Google's mission, "to organize the World's information and make it universally accessible and useful," it is entirely appropriate that they should undertake such an ambitious endeavor of digitizing the World's printed books. While this study identified many of the shortcomings in image quality for works related to

geology, it was found that the vast majority of page images have no scanning errors. The *HathiTrust* Digital Library and Google Book are providing easy access to many works that would otherwise be very difficult to utilize for many researchers. Perhaps more importantly, the *HathiTrust* Digital Library intends to provide long term stewardship and digital access to works in the public domain and has demonstrated a commitment to quality control—as they have already corrected most the errors identified by this project when they were provided with the information on where the errors were.

References Cited

Bailey, R.G., 1978, Description of the ecoregions of the United States: Ogden, Utah, Dept. of Agriculture, Forest Service, 77 p. Available from: http://hdl.handle.net/2027/umn.319510028429007 (September 29, 2010).

Band, J., 2009, The Long and Winding Road to the Google: The John Marshall Review of Intellectual Property Law, v. 9, no. 227, p. 227-339. Available from: http://www.jmripl.com/Publications/Vol9/Issue2/Band.pdf (June 21, 2011).

Darton, R., 2009, Google & the Future of Books: The New York Review of Books, v. 56, no. 2. Available from: http://www.fulminiesaette.it/_uploads/biblioteca/Darnton%20-%20Biblioteca%20Universale%20e%20Google.pdf (June 21, 2011).

Dougherty, W.C., 2010, The Google Books Project: Will it Make Libraries Obsolete?: Journal of Academic Librarianship, v. 36, no. 1, p. 86-89.

Duguid, P., 2007, Inheritance and loss? A brief survey of Google Books: First Monday, v. 12, no. 8. Available from: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/1972/1847 (June 21, 2011) .

HathiTrust, 2010, HathiTrust: A Shared Digital Repository: http://www.hathitrust.org/ (October 29, 2010).

James, R., 2010, An Assessment of the Legibility of Google Books: Journal of Access Services, v. 7, no. 4, p. 223-228.

Jones, E., 2010, Google Books as a general research collection: Library Resources & Technical Services, v. 54, p. 77-89. Available from:

http://national.academia.edu/EdJones/Papers/117683/Google_Books_as_a_General_Research_Collection (June 21, 2011).

Nunberg, G., 2009, Google's book search: A disaster for scholars: The Chronicle Review. Avalible from: http://chronicle.com/article/Googles-Book-Search-A/48245/ (June 21, 2011).

Okerson, A., 2009, The Continuing Saga of the Google Book Settlement: Against the Grain v. 23, no. 3, p. 1-17. Available from: http://www.against-the-grain.com/2010/07/v-22-3-table-of-contents/ (June 21, 2011).

Proskine, E.A., 2006, Google's Technicolor Dreamcoat: A Copyright Analysis of the Google Book Search Library Project: Berkeley Technology Law Journal, v. 21, p. 213. Available from: http://heinonline.org/HOL/LandingPage?collection=journals&handle=hein.journals/berktech21&div=25&id=&page= (June 21, 2011).

Townsend, R.B., 2007, Google Books: Is it good for history?: Perspectives v. 45, no. 6. Available from: http://www.historians.org/perspectives/issues/2007/0709/0709vie1.cfm (June 21, 2011).

York, J., 2010, Personal Electronic Communication with Author, (September 9, 2010).