Structured Abstract:

Purpose – This article reports on a quantitative study of massive digital library (MDL) Google Books’ coverage of Hawaiian and Pacific Books.

Design/methodology/approach – A total of 1,500 books were randomly selected from the University of Hawai‘i at Mānoa’s Hawaiian, Pacific, and general stacks collections. Their level of access was then determined in Google Books by observing whether the books had a metadata record, were full-text searchable, and whether they were available as in snippet, preview, or full-text views.

Findings – Results show that Google Books has a sizable number of metadata records for Hawaiian and Pacific books, but has only a limited number available for full-text searching. In contrast, a larger number of books from the general stacks were available for full-text searching.

Research limitations/implications – Because of the small sample size, margins of error remain large. The field would benefit from a larger size of collection sample. The scope of the project is also limited to Google Books and does not investigate other book digitization projects.

Practical implications – Diversity in librarianship is a major concern for libraries both within the United States, as in the case of historically underrepresented groups as well as in non-English-speaking countries.

Social implications – Diversity in librarianship also concerns the central mission of libraries to provide the basic human right of access to information. Digital libraries must be held to the same standards.

Originality/value – Massive digital libraries such as Google Books need to be more carefully examined; this study contributes to this need.

Keywords: Massive digital libraries (MDL); Google Books; Hawaii; Pacific; historically underrepresented groups; Universal Declaration of Human Rights (UDHR)

Article Type: Research paper

Introduction & Overview:

This study grew out of a question posed by the authors: Is Google Books truly a “Universal Library” or is it an “American Universal Library”? On deeper consideration, this question is not as facetious as it might appear on the surface. As Google Books is now one of the largest digital libraries in the world, serious questions concerning its scope, depth, and breadth should be considered important areas of research.
Many of the libraries contributing items to the Google Books project are located in the 48 contiguous states of the United States of America. Often referred to as “the Google 5,” these libraries include Harvard, Michigan, Stanford, the University of Oxford, and New York Public Library (Lavoie, et al. 2005). Despite the regional and multicultural diversity of the United States, many specialized collections lie outside this geographical constraint, including regions in Europe, South America, Asia and the Pacific. Lavoie et al. demonstrate that English language books represent 49% of the Google Books collection in comparison to 52% of OCLC’s World Cat collection, while German (10%), French (8%), Spanish (5%), Chinese (4%), Russian (4%), and Italian (3%) round out the remainder of the largest non-English books (Lavoie, et al. 2005). Although over 430 languages are represented in total, more than 420 languages are represented in the remaining 16% of books, and only two of these, Japanese and Hebrew, reach 2% of the collection respectively. On the basis of their findings, Lavoie et al. conclude that Google Books may be “more culturally diverse than originally anticipated” (Lavoie et al. 2005).

However, this conclusion of cultural diversity does not satisfy all constituents. European scholars have been the most vocal in criticizing Google Books for failing to include materials with diverse cultural perspectives. In one notable critique, Jean-Noel Jeanneney, President of the Bibliothèque nationale de France, argues that although the Google Books digitization project will result in the digitization of roughly 15 million books, this represents only 15% of Western literature (Bearman 2006). Jeanneney further observes that Google Books, which began as “an American project...that reached out to the British,” (Jeanneney 2006, p. 4), contains an inherent English-language bias stemming from the project’s collection base sources. Jeanneney further critiques Google Books by suggesting that Google’s keyword search provides results that contradict the rankings scholars have provided these works, thereby disregarding well-established cultural values (Bearman 2006). These concerns that Google Books may be biased toward adding English-language materials to its collection deserve further quantitative investigation.

More current studies have focused on some of the flaws in access within the Google Books model. James’s findings in his article, “An Assessment of the Legibility of Google Books,” point toward numerous digitization problems in the project. Nunberg has found metadata errors in the project and believes it to be a flawed system (2009a) (2009b). James and Weiss in their 2012 study have shown that 36% of sampled books contain metadata errors (2012). Pope and Holley demonstrate that many public domain books, despite being digitized and part of the public domain are inaccessible, locked behind Google Books’ limited-, snippet-, and non-views (2011). Most recently, Chen examines Google Books’ limited retrieval of WorldCat books is comprehensive, access issues remain prevalent (2012). Less than 10 percent of records sampled in his study have free full views, 15 percent are visible using the snippet view and 15 percent show previews only (Chen 2012).

Of course, Google Books is not the only massive online digital library project. Other projects such as the Hathi Trust and the Internet Archive are working in a similar manner with various partners for digitization (sources?). However, despite their similarities in size and ambition, comparisons between Google Books and the Internet Archive are not easy. One is part
of a well-funded, for-profit company, the other is a non-profit organization. Some studies have shown overlaps in their approaches. Blakely, for example, shows how both Google Books and the Internet Archive are able to meet specific searching needs for government documents (2009). Hathí Trust is built in part upon the digitization efforts of the Google Books project itself, but includes other library digitization projects as well (Booth). Their current statistics indicate that nearly 10 million volumes have been digitized and included in their project (Hathí).

In light of this recent scholarship showing limitations to the Google Books model, it is important to return to the issue of access and cultural comprehensiveness that Jeanneney and others touched upon in their initial reactions to the project six years earlier.

In the monograph Beyond Article 19, Stephen Edwards argues that libraries should be seen “not as static repositories but as organic and dynamic agents of culture” (Edwards 2010, p. 12). Additionally, Knuth notes that “books animate societies, and libraries collect the stories that give shape and meaning to our lives” (Knuth 2003, p. 2). In his discussion on the Universal Declaration of Human Rights (UDHR) Albarillo argues that when considering issues of access to information one should “move away from our past history of an English-centered library practice” (Albarillo 2010, p. 99). He further argues for an increase in cross-cultural and language training for libraries (Albarillo 2010, p. 104). Seen from these perspectives, Google Books and the other massive digital libraries morph from being merely a vessels carrying culture to a mirror reflecting the culture of the creators and their users, and if they underrepresent certain groups, they could be potentially seen as perpetuating bias.

It is absurd at this juncture, however, to accuse Google Books and other massive online digital libraries of purposely violating any particular group’s human rights or of purposely creating bias. Google Books and the other digital projects are too new and still growing. The ideas of an MDL and a “universal library” are still evolving. Our study is meant to provide a clear indication of where gaps may lie within the Google Books project.

The investigators in this study have taken these growing cross-cultural concerns of diversity and applied them to historically under-represented groups in the United States. As a result, in the case of Hawaiian and Pacific materials, the question of Google Books’ universality might be stated a little differently. Going back to the original question of whether Google Books is a universal library or even an American universal library as Jeanneney might claim, the question the researchers ask is whether Google Books is a “mainland American universal library”? In many ways the question is one about prioritization, canonization and reaching traditionally underserved groups. If values are created by the act of inclusion as well as exclusion, then it would be worthwhile seeing how closely Google Books reaches the ideal of universality.

The University of Hawai’i at Mānoa Libraries have a long history of prioritizing materials relating to Hawaiian and Pacific cultures. According to Stu Dawrs, Senior Pacific Specialist Librarian at UH’s Hamilton Library, UH-Mānoa librarians have spent decades developing these collections. The result of their effort has been a collection of over 150,000 items in the Hawaiian collection, and a Pacific collection of over 130,000. This is one of the largest and most comprehensive collections relating to these subjects in the world (Dawrs 2006). It also represents an important starting point for assessing Google Books’ universality and
comprehensiveness. It should be acknowledged that although the University of Hawaii at Manoa has a few archival materials online, it is not currently working in an official capacity with the Google Books project.

Methodology:

A random sample of 500 books was taken from three collections each, the Hawaiian Collection, the Pacific Collection, and the general stacks collection, for a total of 1,500 books. Selected item records were then searched in Google Books. The primary metadata fields used for searching were title, author, and publication date. Metadata related to page number, publisher, edition and reprint were also collected for identifying book records in Google Books. Matching item records found in Google Books were then evaluated for their level of access. The categories used were ‘Record,’ ‘Snippet,’ ‘Preview,’ and ‘Full’. The ‘Record’ category signifies that Google Books only provides access to a metadata record for the selected item. ‘Snippet’ provides the ability to search the text of the item but displays only a tiny portion of the text. ‘Preview’ also allows for full text searching but allows for only a larger, though still limited, number of pages to be viewed. ‘Full’ offers full-text searching and access to the entire item for unrestricted viewing.

Selected items records were recorded at their highest level of access, and only categorized once. Additionally, selected item records that could not be found in Google Books were recorded in the “no record” category. There are no inter-coder statistics available because all coding and categorizing of the data were done by one of the investigators.

Results:
Figure 1: Graph showing number of items in each collection sampled and their corresponding availability in Google Books.

- Of the 500 randomly sampled Hawaiian collection books, 131 had no record in Google Books, 317 had a metadata record only, 35 had a snippet view available, 11 were available for preview, and 6 could be fully viewed.
- For the Pacific collection books, of the 500 randomly sampled books 116 had no record, 261 had a metadata record only, 98 had a snippet view, 17 were available for preview, and 8 could be fully viewed.
- For the general stacks collection, out of 500 randomly sampled books 40 had no record, 153 had a metadata record only, 229 had a snippet view, 66 were available for preview, and 12 could be fully viewed.
- Additionally, 6 Hawaiian, 6 Pacific collection and 29 general stacks collection books were available for purchase through Google Books.
- Of the 21 total books in the subset of sampled Hawaiian collection books that likely fall within the public domain, 4 of these had no record, 11 had only a metadata record, none (0) had a snippet view and 6 had a full view (29% of titles).
- Of the 24 total books in the subset of sampled Pacific collection books that likely fall within the public domain, 4 of these had no record, 11 had only a metadata record, 3 had a snippet view, and 6 had a full view (25% of titles).
- Of the 18 books in the subset of sampled general stacks collection books that likely fall within the public domain, 3 of these had no record, 3 had only a metadata record, 1 had a snippet view, and 11 had a full view (61% of titles).
- None of the collections had any books available for preview from this subset of the sample.

<table>
<thead>
<tr>
<th>Collection</th>
<th>No Record</th>
<th>Record</th>
<th>Snippet</th>
<th>Full</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hawaiian</td>
<td>4</td>
<td>11</td>
<td>0</td>
<td>6</td>
<td>21</td>
</tr>
<tr>
<td>Pacific</td>
<td>4</td>
<td>11</td>
<td>3</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>General</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>11</td>
<td>18</td>
</tr>
</tbody>
</table>

Figure 2: Table showing number of public domain items in each collection sampled and their corresponding availability in Google Books.

Discussion & Conclusions:

Implications of the statistics:

The results of this research suggest that Google Books’ approach to the mass digitization of books does not provide the adequate coverage of diversity from multiple cultural perspectives.

In the sample of books taken, the general trend shows that there are a greater number of items that either have no record in Google Books or have only metadata records, suggesting that these books have not been digitized. Furthermore, a larger number of books with snippet views exist for a general collection versus the Hawaii and Pacific collections. In fact, there exists
a 2.34 to 1 ratio between the books with snippet views in the General collection and those in the Pacific collection; this increases to 6.54 to 1 between the general stacks and the Hawaiian collection. Books with a preview are also tilted more toward the general stacks collection as well, with a 3.88 to 1 ratio between general stacks books and the Pacific collection and a 6.00 to 1 ratio between general stacks and the Hawaiian collection. Finally, the number of books with full views becomes a little more evenly represented with the ratios flattening to 1.5 to 1 for stacks and Pacific books and 2.0 to 1 for stacks and Hawaiian books.

As far as works in the public domain are concerned, one might expect more evenly matched numbers. This is not the case, however. Despite the possibility of open access and the lack of copyright restrictions, the ratio of books available for full view in the general stacks is still nearly twice that of both the Hawaiian and Pacific books (1.83 to 1). Indeed, when looking at the number of books with metadata records only and those with no records at all, it is obvious that more than half of the books have not been digitized. A similar amount of gaps in coverage were also shown in Lavoie et al.’s study for Japanese language books, which showed that only half of the books available in OCLC’s World Cat had actually been digitized (Lavoie, et al. 2005). The results point toward a large majority of the books in the Hawaiian and Pacific collections not being digitized at all. Metadata records exist for many of these items, but with no views, and in some cases no records of any kind available, one must assume that digitization has not taken place.

After looking at these results, we can see that despite a collection of over 15 million books (just 15% of the estimated 100 million books in Western culture), when looking at the nature and quality of the Google Books corpus, there is much to be desired compared to the Hawaiian and Pacific collections of a library specializing in these subjects.

Bias & Under-represented groups:

This study of Hawaiian and Pacific materials is intended to be a starting point for further research into Google Books’ and other massive digital libraries’ coverage of culturally diverse library materials. As Jeanneney suggests, and the data in this study back him up, there is a real tendency in Google Books to present a culturally biased version of a digital library. At the very least, awareness of this bias may help to ameliorate any negative effects and can provide direction for further digital library development.

Google Books’ initial effort to focus on mainland libraries as partners does appear to have skewed the coverage of their collection toward mainland American user needs. This result is not surprising by itself, as the priorities of mainland academic libraries are by nature not the same as academic libraries in Hawaii. It is not to be argued that they should be, either. Collections developed to meet the needs of local communities have always been the distinguishing marks of every library. However, the issue does raise concerns regarding Google Books’ coverage of traditionally underserved populations, including native Hawaiians and other Pacific Islanders.

Going deeper, one might expect to find other biases reflecting the character of American libraries and the culture and time in which they were created. When Google Books draws upon certain American libraries for their source material, it also incorporates their underlying collection development agendas into its own aggregated collection as a result.
With this understanding, we might ask, then, when will Hawaiian and Pacific materials be given significant coverage in the collection? This is a question of policy and procedure, but it cannot be separated entirely from the question of political and cultural representation, as the European critics of *Google Books* have rightly pointed out (Bearman 2006) (Jeanneney 2006).

**Collection development & the dream of a universal library**

*Google Books’* stated objective to be a “universal library” brings up questions about what their collection development policy might be and what steps they are taking to achieve their goals. There is a risk that the lofty goal of being the “universal library” would lead to having a collection development policy that is functionally and practically equivalent to having no policy at all. The desire to acquire everything, without qualification, leaves happenstance as the primary force in collection development.

Whereas the objective construction of a universal library gives fair and sufficient representation to all cultural perspectives, subjective construction focuses on one culture, and marginalizes others both intentionally and unintentionally. As *Google Books*, unlike other mass-digitization projects, is a for-profit corporation relying on agreements with primarily mainland US academic and public libraries to scan books, the biases of these libraries as well as the for-profit model could be incorporated into the digitization project, creating a subjective “universal library” that marginalizes materials not well-represented in their sources.

It is likely that any massive digital library project that aggregates records from multiple collections, including the Hathi Trust, Internet Archive and all others of similar ilk, will be vulnerable to the limitations of its source material. If researchers are able to pinpoint the weaknesses of collections then at least the illusion of universality when none exists can be challenged. To avoid such pitfalls of coverage, perhaps an approach different to *Google Books’* massive digital library development is warranted.

**Limitations of the study:**

The sample size of this study was quite small; a larger sampling size might improve the accuracy of the results. Furthermore, the study’s relatively narrow scope did not take into account the coverage of digitized works of Hawaiian/Pacific works in other massive online digital libraries such as the Internet Archive or the Hathi Trust. Further, the University of Hawai‘i is not partnering currently with large-scale digitization projects. Should UH be involved in such a project, findings might be altered.

**Further Research:**

Though the sampled collections were small, the results of this study point toward fruitful areas of research. In particular the results of this study could be used in comparison with investigations of the digital collection development of traditionally underrepresented groups within the United States, Europe, Japan (especially Okinawan, Burakumin, Ainu and Zainichi Korean cultures), and other countries.
Additionally, the researchers look to develop a wider and more comprehensive investigation of the coverage of Hawaiian and Pacific materials by examining other massive online digital libraries, including the Hathi Trust, the Internet Archive and others.

Finally, another area of investigation would involve examining the publishers working with Google Books and confirming whether there is a correlation with publishers' copyright and open-access policies and the ultimate choice to be digitized. Looking at impact of orphan works on digitization choices will also provide fruitful areas of investigation. The results showed that a disproportionate amount of books in the public domain had not been digitized. Investigating the reasons behind this gap would also provide interesting insight into improving mass-digitization efforts.

References:


Dawrs, S. (2006), "University of Hawai‘i Hamilton Library Department History, Special Collections."


