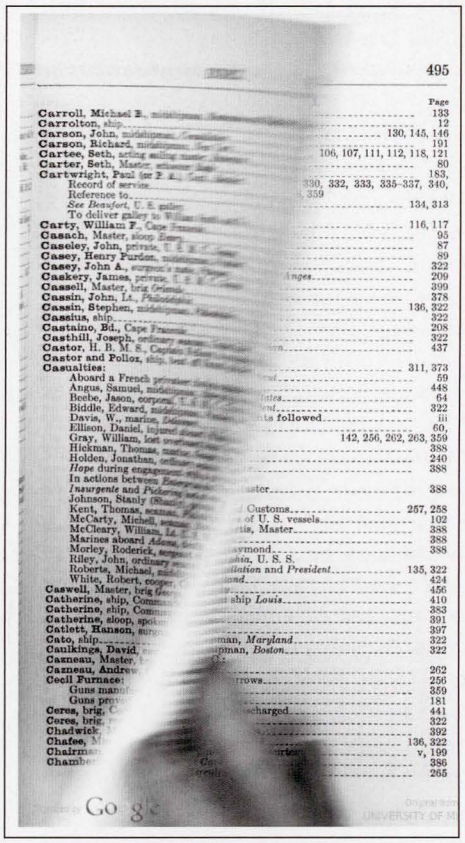
Using Google Books, HathiTrust, and other Online Book Sites

Google's goal to digitize the world's libraries is one of its typical "shoot-for-the-moon" projects, which seem nearly impossible until they turn out to be well on their way to reaching the goal. Its impact is already immeasurable, and will continue to grow. Nevertheless, the project is not without its challenges, and there is much one should know about the process and alternatives.

Begun in secret in 2002 and launched publically in 2004, the concept was something that Google's cofounders always planned to pursue. The project had both altruistic and commercial aims: it has always had a goal of making out-of-print resources available to anyone, but also has involved direct work with publishers to earn revenue when making publishers' content findable, if not accessible. The addition of library partners has greatly expanded the number of out-of-copyright or out-of-print books (and magazines!) findable in **Google Books**.

Results from Google Books will appear in a standard Google search, and you can separate them out by clicking on "Books," between the search box and the results. (Sometimes, "Books" appears on the dropdown "More" menu.) You can also start your search at **https://books.google.com** to limit the initial search to books. There are two types of results: snippet view and full view. In the latter case, you can see the full page of the book in question, with your search results highlighted on the page. The entire book may be visible, or a few pages here and there will be missing. This is for copyright reasons, not because Google skipped some pages. In snippet view, Google has scanned the entire book but can only show small portions of the text because copyright owners have required that they not share the complete text, or even large portions of it.



Google Books is not without problems. The OCR (Optical Character Recognition) text conversion has been problematic, at times. Try searching for "modem" in 19th-century titles, for instance—these will nearly always be a poorly interpreted version of the word "modern." In addition, the metadata—the information describing each title—is sometimes incorrect, making it hard to find some items, particularly volumes in a series, like the Navy Records Society volumes. Google uses high-volume, large-scale workflows, and correcting errors here and there is generally not worth their time. Occasionally, the scans will show pages in the process of being turned, like this example I found last year.

Other digitization projects certainly exist. **Project Gutenberg** (**https://www.gutenberg.org**) was started in 1971 (really!); text was manually keyed in until 1989, when scanners began being used. Each title is available in multiple formats, and nearly all of their 53,000 titles are out of copyright. **HathiTrust** (**https://www.hathitrust.org**) is a very large compiler of digital titles from many sources, including Google Books, the Internet Archive, and Microsoft. HathiTrust has a number of particularly valuable features: you can choose to either search the full text of the books available, or search the catalog that describes the books. If you want to look for a broad topic, like "Australian schooners," you might try searching the catalog to locate books on the subject. If you want to look something more specific, then searching the full text for, say, "Wollongong schooners" might be better. You can use AND and OR terms when searching, and * as a wildcard (a search for "sail*" will return "sail," "sails," "sailors," "sailing," etc.), but there are few other advanced search functions. Most importantly, if you can log in to HathiTrust as a member of a partner institution, you can then download PDFs of most titles, or create collections of specific books, across which you can then search within that collection.

In fact, HathiTrust is often a better option than Google Books. Google makes blanket decisions about book access, while HathiTrust has a collection of librarian volunteers working to investigate and determine copyright whenever possible. For example, Google Books will only show snippets of *Naval Documents Related to the Quasi-War*, a multi-volume set of primary documents published by the US Government in the 1930s. Google does not recognize that since the set was published by the US Government, it is thus not protected by copyright. It just sees a date later than 1922, and assumes it's in copyright. HathiTrust, on the other hand, will allow you to view the title, and if you're at one of the more than 100 partner libraries, you can download every volume. (In this case, you can also download these titles from the nascent **American Naval Records Society**, at **http://www.ibiblio.org/anrs/quasi.html**.)

Sometimes, there's nothing like having a print copy of a title. Used booksellers, such as ABE Books (**https://www.abebooks.com**), are often your best bet. **Amazon.com** is a frustrating option, because most options come from vendors who will simply pass on the order to a print-on-demand publisher, who will print a copy of the Google Books version, errors and all. When looking at purchasing options on used-book websites, be careful what you select. Some sellers will include images of the actual copy you're buying, so you'll know it's original, but if the listed publisher is a university or a library, that's metadata taken straight from Google, and you'll almost certainly receive a print-on-demand copy of the Google scan.

Suggestions for other sites worth mentioning are welcome at peter@shipindex.org. See **www.shipindex.org** for a free compilation of over 150,000 ship names from indexes to dozens of books and journals. ⚓