

Google Print vs. Onsite Collections

Don't send your paper copies off to remote storage just yet

by Thomas Mann

When the topic of Google Print came up at a recent meeting, one librarian made an interesting comment. His supervisor, he said, looked forward to having 15 million electronic books so he could send to remote storage every paper copy with an online equivalent. That struck me as unwise, and it ties in with another proposed scheme advocated by some prominent members of our profession—that libraries ought to use shared warehouses, not merely to house little-used older materials, but to avoid “costly duplication” of current purchases.

In this vision, the primary function of a library is to offer access to the internet and commercial databases and, beyond that, to serve mainly as a community social center with meeting rooms, shared study facilities, and coffee bars. But let's consider the virtues of focused browsing in onsite collections, as well as the adequacy of Google's keyword access—either in Google Print itself or in plans to make library catalogs “more like Google.”

Researchers in this country have long been accustomed to having direct access to large book collections shelved in subject-classified stacks. What benefits arise from this configuration, as opposed to having digital texts at hand? What difference is there to scholarship in the way books are stored and presented?

Let me give an example. A graduate student writing on Paul Valéry had a problem with unverified information regarding the French Symbolist poet's connection to the famous Dreyfus affair of the 1890s. The student had heard information from Valéry's relatives that he had signed some kind of petition on the issue, but no specifics. The ARTFL database of digitized French texts did not help; neither did biographies, several subscription databases, or two massive bibliographies on Valéry. Finally, I had to go into the Library of Congress stacks, where there are 186 volumes on six shelves in the classes DC354–354.9 (“Dreyfus affair”).

I was looking for one book of primary sources that I located in the online catalog with the subject heading “Dreyfus, Alfred, 1859–1935—Trials, litigation, etc.—Sources”; however, it did not include the Valéry petition. But on the shelf above it I noticed another book (Patrice Bousset, *L'affaire Dreyfus et la presse*, 1960) which, it turned out, did indeed have the necessary information. As an extra serendipitous bonus, the same source contained additional information about one of Valéry's close friends—information that solved another problem the researcher hadn't specifically asked about.

Could she have found this source if the same 186 books had been shelved in separate tubs in a remote warehouse? Could she have found it if all of the texts were digitized and fully keyword searchable? The answer, in either case, is “no.” This may seem counterintuitive to information theorists, but I suspect working scholars will immediately know the reason: Classified shelving enables researchers to recognize sources whose keywords they could never specify in advance.

If the books were stored offsite, they could not have been retrieved as a group. Even if the library were willing to deliver all 186 volumes, the researcher would have had to identify, and fill out a separate request for, each volume individually and then wait weeks for all to be retrieved—at the same time not having any clue in advance which one, if any, might repay such extraordinary effort. That degree of hassle puts such persistence beyond the realm of any behavior that can reasonably be expected of any researchers, even that of senior scholars.

Keyword quandary

By contrast, when the same books on a common subject are shelved immediately adjacent to each other, arranged for in-depth inspection, a researcher can rapidly recognize valuable information, even when it's found on a single page. The difference is not simply that of timeliness of access; it's more a matter of recognition versus prior specification. Researchers who cannot clearly specify in advance all the words they want to see can—and do—nonetheless recognize that information when it is presented immediately in front of them within a manageably segregated group of likely sources.

But what if the books were all fully digitized by Google Print? In this same Paul Valéry example, the researcher in question told me she could not have found the necessary information even in such a huge full-text database. Why not?

The core problem remains that the researcher did not know in advance the right keywords to type in. The relatives had simply told her that Valéry had signed a “petition” (a *liste*) con-

THOMAS MANN is the author of *The Oxford Guide to Library Research* (Oxford, 3d ed., 2005). He has been a reference librarian at the Library of Congress for 24 years. His views are not necessarily those of the LC administration.



nected to the Dreyfus case. It turns out it was actually a subscription fund to provide money for the widow of one of the individuals involved in the scandal. And the French text did not use the words *petition* or *liste* to describe the roster—it used the terms *souscription* and *souscripteurs* instead. In other words, the scholar would not have been able to type the right keywords into the Google search box even if the full text were searchable.

A combined search in Google Web on “Paul Valéry” and “Dreyfus” produces over 3,500 hits. Keyword searching in the proposed Google Print file is likely to produce similar mountains of inadequately sorted chaff. Nor would Google have singled out this one very obscure text based on the frequency of other sites’ links to it. The needed information it contains would be utterly buried, beyond any hope of retrieval, by Google’s inadequate “relevance ranking” search software. Simply having billions more keywords to search is not a solution to such problems; it is a positive exacerbation of them.

Book collections that are categorized by subject on library shelves are atomized like colored mosaics in Google Print. A search can retrieve all the red or the blue or the yellow stones in separate piles, or all of them together, but not in a way that shows them related to each other in a coherent picture. For example, a keyword search for “death penalty” will not recover titles using such terms as *capital punishment*, *death row*, *legal execution*, *Todesstraffe*, or *peine de mort*.

Google’s software can only manipulate results *within* each keyword-defined set; it cannot build bridges *among* multiple sets using different words for the same idea, or covering different aspects of the same subject. Even within any given set, relevance ranking cannot segregate the right words into the right conceptual contexts while simultaneously filtering out appearances of the same words in the wrong contexts. Such problems are large enough to destroy the very possibility of systematic subject searching in research collections: If keyword searches are all we can do—no matter how the results are displayed—it will no longer be possible to get coherent overviews of the books relevant to a topic, nor to segregate the few relevant books that are found from all of the irrelevant ones that happen to contain the same words.

The zen of serendipity

The Valéry example is not merely anecdotal. Studies of information-seeking behavior verify that focused browsing in library bookstacks—enabling scholars to simply *recognize* the information that they cannot specify in advance—continues to be regarded as essential to substantive scholarship.

■ A 2004 University of Oklahoma study noted that “the importance of serendipitous browsing in library collections cannot be overemphasized.”

■ The *College Student Experiences Questionnaire*, with

data representing responses from more than 300,000 students between 1984 and 2002, found that 65.5% of male students and 63.2% of female students reported that they “found something interesting [through] browsing” either “occasionally,” “often,” or “very often”; for students in “Doctoral Intensive” programs the overall figure was 67.7%. (Percentages like these, in presidential elections, would be designated landslides.)

■ A 2003 survey by Margaret Stieg Dalton and Laurie Charnigo on “Historians and Their Information Sources” reports that “informal means of discovery like book reviews and browsing remain important, as does the need for comprehensive searches.” (This, by the way, gives the lie to assertions that researchers today are “wired differently” from their predecessors and “we better get used to it.”)

It is especially the latter—comprehensive searches—that would no longer be possible in Google Print. The admirable goal of freeing book texts from confinement within library walls would entail the unintended consequence of destroying systematic subject access to them. It is not a good trade-off if all the books are freely available but researchers can no longer find the majority of relevant ones or see the forest for the trees.

Browsing in the stacks, of course, is something that researchers usually do themselves; it’s only in a closed stacks library, such as the Library of Congress, where reference librarians with stacks access are called upon to find information that can’t be found in any way other than focused browsing. I’ve found information, through focused browsing, on traveling libraries circulated among lighthouses at the turn of the last century. Within the 438 volumes shelved in the class area VK1000–1025 (“Lighthouse service”) I found 15 books that had directly relevant sections. In comparison, an advanced Google search, combining “libraries” with either “lighthouse” or “lighthouses,” produces 221,000 hits. Google’s software is incapable of segregating the right words into only the right context.

Is the prospect of “getting something quickly” on the internet so attractive that we are willing to sacrifice time-tested mechanisms that make scholarship—not just information seeking—possible in our research libraries’ collections? Every faculty library committee in every college and university needs to carefully monitor the direction in which its library administration is moving.

The same books that are arranged in onsite, subject-classified bookshelves become, suddenly and radically, much less discoverable when they are stored in offsite warehouses or replaced by digitized texts searchable only via keyword “relevance ranking” mechanisms. Our profession is in the grip of an uncritical infatuation with keyword searching as the sole avenue of access to book collections; if this is not corrected and counterbalanced, scholars throughout the nation may soon find that we librarians have traded our birthright for a mess of pottage. ❖

Focused browsing in library bookstacks is still essential to substantive scholarship.