Dan Cohen's Digital Humanities Blog » Blog Archive » Is Google Good For History?

[*These are my prepared remarks for a talk I gave at the American Historical Association Annual Meeting, on January 7, 2010, in San Diego. The panel was entitled "Is Google Good for History?" and also featured talks by Paul Duguid of the University of California, Berkeley and Brandon Badger of Google Books. Given my propensity to go rogue, what I actually said likely differed from this text, but it represents my fullest, and, I hope, most evenhanded analysis of Google.*]

Is Google good for history? Of course it is. We historians are searchers and sifters of evidence. Google is probably the most powerful tool in human history for doing just that. It has constructed a deceptively simple way to scan billions of documents instantaneously, and it has spent hundreds of millions of dollars of its own money to allow us to read millions of books in our pajamas. Good? How about Great?

But then we historians, like other humanities scholars, are natural-born critics. We can find fault with virtually anything. And this disposition is unsurprisingly exacerbated when a large company, consisting mostly of better-paid graduates from the other side of campus, muscles into our turf. Had Google spent hundreds of millions of dollars to build the Widener Library at Harvard, surely we would have complained about all those steps up to the front entrance.

Partly out of fear and partly out of envy, it's easy to take shots at Google. While it seems that an obsessive book about Google comes out every other week, where are the volumes of criticism of ProQuest or Elsevier or other large information companies that serve the academic market in troubling ways? These companies, which also provide search services and digital scans, charge universities exorbitant amounts for the privilege of access. They leech money out of library budgets every year that could

be going to other, more productive uses.

Google, on the other hand, has given us Google Scholar, Google Books, newspaper archives, and more, often besting commercial offerings while being freely accessible. In this bigger picture, away from the myopic obsession with the Biggest Tech Company of the Moment (remember similar diatribes against IBM, Microsoft?), Google has been very good for history and historians, and one can only hope that they continue to exert pressure on those who provide costly alternatives.

Of course, like many others who feel a special bond with books and our cultural heritage, I wish that the Google Books project was not under the control of a private entity. For years I have called for a public project, or at least a university consortium, to scan books on the scale Google is attempting. I'm envious of France's recent announcement to spend a billion dollars on public scanning. In addition, the Center for History and New Media has a strong relationship with the Internet Archive to put content in a non-profit environment that will maximize its utility and distribution and make that content truly free, in all senses of the word. I would much rather see Google's books at the Internet Archive or the Library of Congress. There is some hope that HathiTrust will be this non-Google champion, but they are still relying mostly on Google's scans. The likelihood of a publicly funded scanning project in the age of Tea Party reactionaries is slim.

* * *

Long-time readers of my blog know that I have not pulled punches when it comes to Google. To this day the biggest spike in readership on my blog was when, very early in Google's book scanning project, I casually posted a scan of a human hand[1] I found while looking at an edition of Plato. The post ended up on Digg, and since then it has been one of the many examples used by Google's detractors to show a lack of quality in their library project.

Let's discuss the quality issues for a moment, since it is one point of

obsession within the academy, an obsession I feel is slightly misplaced. Of course Google has some poor scans—as the saying goes, haste makes waste—but I've yet to see a *scientific* survey of the overall percentage of pages that are unreadable or missing (surely a miniscule fraction in my viewing of scores of Victorian books). Regarding metadata errors, as Jon Orwant of Google Books has noted[2], when you are dealing with a trillion pieces of metadata, you are likely to have millions of errors in need of correction. Let us also not pretend the bibliographical world beyond Google is perfect. Many of the metadata problems with Google Books come from library partners and others outside of Google.

Moreover, Google likely has remedies for many of these inadequacies. Google is constantly improving its OCR and metadata correction capabilities, often in clever ways. For instance, it recently acquired the reCAPTCHA system from Carnegie Mellon, which uses unwitting humans who are logging into online services to transcribe particularly hard or smudged words from old books. They have added a feedback mechanism for users to report poor scans. Truly bad books can be rescanned or replaced by other libraries' versions. I find myself nonplussed by quality complaints about Google Books that have engineering solutions. That's what Google does; it solves engineering problems very well.

Indeed, we should recognize (and not without criticism, as I will note momentarily) that at its heart, Google Books is the outcome, like so many things at Google, of a engineering challenge and a series of mathematical problems: How can you scan tens of million books in a decade? It's easy to say they should do a better job and get all the details right, but if you do the calculations with those key variables, as I assume Brandon and his team have done, you'll probably see that getting a nearly perfect library scanning project would take a hundred years rather than ten. (That might be a perfectly fine trade-off, but that's a different argument or a different project.) As in OCR, getting from 99% to 99.9% accuracy would probably take an order of magnitude longer and be an

order of magnitude more expensive. That's the trade-off they have decided to make, and as a company interested in search, where near-100% accuracy is unnecessary, and considering the possibilities for iterating toward perfection from an imperfect first version, it must have been an easy decision to make.

* * *

Google Books is incredibly useful, even with the flaws. Although I was trained at places with large research libraries of Google Books scale, I'm now at an institution that is far more typical of higher ed, with a mere million volumes and few rare works. At places like Mason, Google Books is a savior, enabling research that could once only be done if you got into the right places. I regularly have students discover new topics to study and write about through searches on Google Books. You can only imagine how historical researchers and all students and scholars feel in even less privileged places. Despite its flaws, it will be the the source of much historical scholarship, from around the globe, over the coming decades. It is a tremendous leveler of access to historical resources.

Google is also good for history in that it challenges age-old assumptions about the way we have done history. Before the dawn of massive digitization projects and their equally important indices, we necessarily had to pick and choose from a sea of analog documents. All of that searching and sifting we did, and the particular documents and evidence we chose to write on, were—let's admit it—prone to many errors. Read it all, we were told in graduate school. But who ever does? We sift through large archives based on intuition; occasionally we even find important evidence by sheer luck. We have sometimes made mountains out of molehills because, well, we only have time to sift through molehills, not mountains. Regardless of our technique, we always leave something out; in an analog world we have rarely been comprehensive.

This widespread problem of anecdotal history, as I have called it, will only get worse. As more documents are scanned and go online, many

works of historical scholarship will be exposed as flimsy and haphazard. The existence of modern search technology should push us to improve historical research. It should tell us that our analog, necessarily partial methods have had hidden from us the potential of taking a more comprehensive view, aided by less capricious retrieval mechanisms which, despite what detractors might say, are often more objective than leafing rapidly through paper folios on a time-delimited jaunt to an archive.

In addition, listening to Google may open up new avenues of exploring the past. In my book[3] *Equations from God* I argued that mathematics was generally considered a divine language in 1800 but was "secularized" in the nineteenth century. Part of my evidence was that mathematical treatises, which often contained religious language in the early nineteenth century, lost such language by the end of the century. By necessity, researching in the pre-Google Books era, my textual evidence was limited—I could only read a certain number of treatises and chose to focus (I'm sure this will sound familiar) on the writings of high-profile mathematicians. The vastness of Google Books for the first time presents the opportunity to do a more comprehensive scan of Victorian mathematical writing for evidence of religious language. This holds true for many historical research projects.

So Google has provided us not only with free research riches but also with a helpful direct challenge to our research methods, for which we should be grateful. Is Google good for history? Of course it is.

* * *

But does that mean that we cannot provide constructive criticism of Google, to make it the best it can be, especially for historians? Of course not. I would like to focus on one serious issue that ripples through many parts of Google Books.

For a company that is a champion of openness, Google remains strangely closed when it comes to Google Books. Google Books seems to operate in

ways that are very different from other Google properties, where Google aims to give it all away. For instance, I cannot understand why Google doesn't make it easier for historians such as myself, who want to do technical analyses of historical books, to download them en masse more easily. If it wanted to, Google could make a portal to download all public domain books tomorrow. I've heard the excuses from Googlers: But we've spent millions to digitize these books! We're not going to just give them away! Well, Google has also spent millions on software projects such as Android, Wave, Chrome OS, and the Chrome browser, and they are giving those away. Google's hesitance with regard to its books project shows that openness goes only so far at Google. I suppose we should understand that; Google is a company, not public library. But that's not the philanthropic aura they cast around Google Books at its inception or even today, in dramatic op-eds touting the social benefit of Google Books.

In short, complaining about the quality of Google's scans distracts us from a much larger problem with Google Books. The real problem—especially for those in the digital humanities but increasingly for many others—is that Google Books is only open in the read-a-book-in-my-pajamas way. To be sure, you can download PDFs of many public domain books. But they make it difficult to download the OCRed text from multiple public domain books–what you would need for more sophisticated historical research. And when we move beyond the public domain, Google has pushed for a troubling, restrictive regime for millions of so-called "orphan" books.

I would like to see a settlement that offers greater, not lesser access to those works, in addition to greater availability of what Cliff Lynch[4] has called "computational access" to Google Books, a higher level of access that is less about reading a page image on your computer than applying digital tools to many pages or books at one time to create new knowledge and understanding. This is partially promised in the Google Books settlement, in the form of text-mining research centers, but those centers will be behind a velvet rope and I suspect the casual historian will be

unlikely to ever use them. Google has elaborate APIs, or application programming interfaces, for most of its services, yet only the most superficial access to Google Books.

For a company that thrives on openness and the empowerment of users and software developers, Google Books is a puzzlement. With much fanfare, Google has recently launched—evidently out of internal agitation—what it calls a "Data Liberation Front," to ensure portability of data and openness throughout Google. On dataliberation.org, the website for the front, these Googlers list 25 Google projects and how to maximize their portability and openness—virtually all of the main services at Google. Sadly, Google Books is nowhere to be seen, even though it also includes user-created data, such as the My Library feature, not to mention all of the data—that is, books—that we have all paid for with our tax dollars and tuition. So while the Che Guevaras put up their revolutionary fist on one side of the Googleplex, their colleagues on the other side are working with a circumscribed group of authors and publishers to place messy restrictions onto large swaths of our cultural heritage through a settlement that few in the academy support.

Jon Orwant and Dan Clancy and Brandon Badger have done an admirable job explaining much of the internal process of Google Books. But it still feels removed and alien in way that other Google efforts are not. That is partly because they are lawyered up, and thus hamstrung from responding to some questions academics have, or from instituting more liberal policies and features. The same chutzpah that would lead a company to digitize entire libraries also led it to go too far with in-copyright books, leading to a breakdown with authors and publishers and the flawed settlement we have in front of us today.

We should remember that the reason we are in a settlement now is that Google didn't have enough chutzpah to take the higher, tougher road—a direct challenge in the courts, the court of public opinion, or the Congress to the intellectual property regime that governs many books and makes them difficult to bring online, even though their authors and

publishers are long gone. While Google regularly uses its power to alter markets radically, it has been uncharacteristically meek in attacking head-on this intellectual property tower and its powerful corporate defenders. Had Google taken a stronger stance, historians would have likely been fully behind their efforts, since we too face the annoyances that unbalanced copyright law places on our pedagogical and scholarly use of textual, visual, audio, and video evidence.

I would much rather have historians and Google to work together. While Google as a research tool challenges our traditional historical methods, historians may very well have the ability to challenge and make better what Google does. Historical and humanistic questions are often at the high end of complexity among the engineering challenges Google faces, similar to and even beyond, for instance, machine translation, and Google engineers might learn a great deal from our scholarly practice. Google's algorithms have been optimized over the last decade to search through the hyperlinked documents of the Web. But those same algorithms falter when faced with the odd challenges of change over centuries and the alienness of the past and old books and documents that historians examine daily.

Because Google Books is the product of engineers, with tremendous talent in computer science but less sense of the history of the book or the book as an object rather than bits, it founders in many respects. Google still has no decent sense of how to rank search results in humanities corpora. Bibliometrics and text mining work poorly on these sources (as opposed to, say, the highly structured scientific papers Google Scholar specializes in). Studying how professional historians rank and sort primary and secondary sources might tell Google a lot, which it could use in turn to help scholars.

Ultimately, the interesting question might not be, Is Google good for history? It might be: Is history good for Google? To both questions, my answer is: Yes.

This entry was posted on Thursday, January 7th, 2010 at 6:00 pm and is filed under Google[5]. You can follow any responses to this entry through the RSS 2.0[6] feed. You can leave a response[7], or trackback[8] from your own site.

References

1. ^ a scan of a human hand (www.dancohen.org)
2. ^ has noted (languagelog.ldc.upenn.edu)
3. ^ my book (www.dancohen.org)
4. ^ Cliff Lynch (www.cni.org)
5. ^ View all posts in Google (www.dancohen.org)
6. ^ RSS 2.0 (www.dancohen.org)
7. ^ leave a response (www.dancohen.org)
8. ^ trackback (www.dancohen.org)