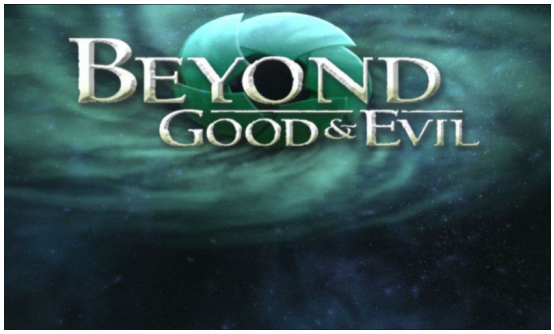
Dan Cohen's Digital Humanities Blog » Blog Archive » Google Books: Champagne Or Sour Grapes?



Is it possible to have a balanced discussion of Google's outrageously ambitious and undoubtedly flawed project to scan tens of millions of books in dozens of research libraries? I have noted in this space the advantages[1] and disadvantages[2] of Google Books[3]—sometimes both at one time[4]. Heck, the only time this blog has ever been seriously "dugg[5]" is when I noted the appearance of fingers in some Google scans[6]. Google Books is an easy target.

This week Paul Duguid has received a lot of positive press (e.g., Peter Brantley[7], if:book[8]) for his dressing down of Google Books, "Inheritance and loss? A brief survey of Google Books[9]." It's a very clever article, using poorly scanned Google copies of Lawrence Sterne's absurdist and raunchy comedy *Tristram Shandy* to reveal the extent of Google's folly and their "disrespect" for physical books.

I thought I would enjoy reading Duguid's article, but I found myself oddly unenthusiastic by the end.

Of course Google has poor scans—as the saying goes, haste makes waste—but this is not a scientific survey of the percentage of pages that are unreadable or missing (surely less than 0.1% in my viewing of scores of Victorian books). Nor does the article note that Google might have possible remedies for some of these inadequacies. For example, they almost certainly have higher-resolution, higher-contrast scans that are different than the lo-res ones they display (a point made at the Million Books workshop[10]; they use the originals for OCR), which they can revisit to produce better copies for the web. Just as they have recently added commentary to Google News[11], they could have users flag

problematic pages. Truly bad books could be rescanned or replaced by other libraries' versions.

Most egregiously, none of the commentaries I have seen on Duguid's jeremiad have noted the telling coda to the article: "This paper is based on a talk given to the Society of Scholarly Publishers, San Francisco, 6 June 2007. I am grateful to the Society for the invitation." The question of playing to the audience obviously arises.

Google Books will never be perfect, or even close. Duguid is right that it disrespects age-old, critical elements of books. (Although his point that Google disrespects metadata strangely fails to note that Google is one of the driving forces behind the Future of Bibliographic Control[12] meetings, which are *all* about metadata.) Google Books is the outcome, like so many things at Google, of a mathematical challenge: How can you scan tens of millions of books in five years? It's easy to say they should do a better job and get all the details right, but if you do the calculations of that assessment, you'll probably see that the perfect library scanning project would take 50 years rather than 5. As in OCR, getting from 98% to 100% 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 (I have seen OCR specialists estimate that even 90% accuracy is perfectly fine for search), it must have been an easy decision to make.

Complaining about the quality, thoroughness, and fidelity of Google's (public) scans distracts us from the larger problem of Google Books. As I have argued repeatedly in this space, the real problem—especially for those in the digital humanities but also for many others—is that Google Books is not open. Recently they have added the ability to view some books in "plain text" (i.e., the OCRed text, but it's hard to copy text from multiple pages at once), and even in some cases to download PDFs of public domain works. But those moves don't go far enough for scholarly needs. We need what Cliff Lynch[13] of CNI[14] has called "computational

access," a higher level of access that is less about reading a page image on your computer than applying digital tools and analyses to many pages or books at one time to create new knowledge and understanding.

An API[15] would be ideal for this purpose if Google doesn't want to expose their entire collection. Google has APIs for most of their other projects—why not Google Books?

[*Image courtesy of Ubisoft*.]

This entry was posted on Thursday, August 16th, 2007 at 9:53 am and is filed under Books[16], Digitization[17], Google[18]. You can follow any responses to this entry through the RSS 2.0[19] feed. You can leave a response[20], or trackback[21] from your own site.

References

1. ^ advantages (www.dancohen.org)
2. ^ disadvantages (www.dancohen.org)
3. ^ Google Books (books.google.com)
4. ^ both at one time (www.dancohen.org)
5. ^ dugg (www.digg.com)
6. ^ the appearance of fingers in some Google scans (www.dancohen.org)
7. ^ Peter Brantley (radar.oreilly.com)
8. ^ if:book (www.futureofthebook.org)
9. ^ Inheritance and loss? A brief survey of Google Books (www.firstmonday.org)
10. ^ the Million Books workshop (www.dancohen.org)
11. ^ recently added commentary to Google News (blogoscoped.com)
12. ^ Future of Bibliographic Control (www.loc.gov)
13. ^ Cliff Lynch (www.cni.org)
14. ^ CNI (www.cni.org)
15. ^ API (www.dancohen.org)
16. ^ View all posts in Books (www.dancohen.org)
17. ^ View all posts in Digitization (www.dancohen.org)

18. ^ View all posts in Google (www.dancohen.org)
19. ^ RSS 2.0 (www.dancohen.org)
20. ^ leave a response (www.dancohen.org)
21. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen's Digital Humanities Blog » Blog Archive » Google Books: Champagne or Sour Grapes?*

http://www.dancohen.org/2007/08/16/google-books-champagne-or-sour-grapes/

Readability — An Arc90 Laboratory Experiment

http://lab.arc90.com/experiments/readability