Dan Cohen's Digital Humanities Blog » Blog Archive » Mass Digitization Of Books: Exit Microsoft, What Next?

So Microsoft[1] has left the business of digitizing millions of books[2]—apparently because they saw it as no business at all[3].

This leaves Microsoft's partner (and our partner on the Zotero[4] project), the Internet Archive[5], somewhat in the lurch, although Microsoft has done the right thing and removed the contractual restrictions[6] on the books they digitized so they may become part of IA's fully open collection[7] (as part of the broader Open Content Alliance[8]), which now has about 400,000 volumes. Also still on the playing field is the Universal Digital Library[9] (a/k/a the Million Books Project), which has 1.5 million volumes.

And then there's Google[10] and its Book Search[11] program. For those keeping score at home, my sources tell me that Google, which coyly likes to say it has digitized "over a million books" so far, has actually finished scanning *five* million. It will be hard for non-profits like IA to catch up with Google without some game-changing funding or major new partnerships.

Foundations like the Alfred P. Sloan Foundation[12] have generously made substantial (million-dollar) grants to add to the digital public domain. But with the cost of digitizing 10 million pre-1923 books at around $300 million, where might this scale of funds and new partners come from? To whom can the Open Content Alliance turn to replace Microsoft?

Frankly, I've never understood why institutions such as Harvard, Yale, and Princeton haven't made a substantial commitment to a project like OCA. Each of these universities has seen its endowment grow into the tens of billions in the last decade, and each has the means and (upon

reflection) the motive to do a mass book digitization project of Google's scale. $300 million sounds like a lot, but it's less than 1% of Harvard's endowment and my guess is that the amount is considerably less than all three universities are spending to build and fund laboratories for cutting-edge sciences like genomics. And a 10 million public-domain book digitization project is just the kind of outrageously grand project HYP should be doing, especially if they value the humanities as much as the sciences.

Moreover, Harvard, Yale, and Princeton find themselves under enormous pressure to spend more of their endowment for a variety of purposes, including tuition remission and the public good. (Full and rather vain disclosure: I have some relationship to all three institutions[13]; I complain because I love.) Congress might even get into the act, mandating that universities like HYP spend a more generous minimum percentage of their endowment every year, just like private foundations who benefit (as does HYP, though in an indirect way) from the federal tax code.

In one stroke HYP could create enormous good will with a moon-shot program to rival Google's: free books for the world. (HYP: note the generous reaction to, and the great press for, MIT's OpenCourseWare program[14].) And beyond access, the project could enable new forms of scholarship through computational access to a massive corpora of full texts.

Alas, Harvard and Princeton partnered with Google long ago. Princeton has committed to digitizing about one million volumes with Google; Harvard's number is unclear, but probably smaller. The terms of the agreement with Google are non-exclusive; Harvard and Princeton could initiate their own digitization projects or form other partnerships. But I suspect that would be politically difficult since the two universities are getting free digitization services from Google and would have to explain to their overseers why they want to replace free with very expensive. (The answer sounds like Abbott and Costello: the free program produces

something that's not free, while the expensive one is free.)

If Google didn't exist, Harvard[15] would probably be the most obvious candidate to pull off the Great Digitization of Widener[16]. Not only does it have the largest endowment; historian Robert Darnton, a leader in thinking about the future (and the past) of the book, is now the director of the Harvard library system. Harvard also recently passed an open access mandate for the publications of its faculty.

Princeton[17] has the highest per-student endowment of any university, and could easily undertake a mass digitization project of this scale. Perhaps some of the many Princeton alumni who went on to vast riches on the Web, such as EBay[18]'s Meg Whitman (who has already given $100 million to Princeton) or Amazon[19]'s Jeff Bezos, could pitch in.

But Harvard's and Princeton's Google "non-exclusive" partnership makes these outcomes unlikely, as does the general resistance in these universities to spending science-scale funds outside of the sciences (unless it's for a building).

That leaves Yale[20]. Yale chose Microsoft[21] last year to do its digitization, and has now been abandoned right in the middle of its project. Since Microsoft is apparently leaving its equipment and workflow in place at partner institutions, Yale could probably pick up the pieces with an injection of funding from its endowment or from targeted alumni gifts. Yale just spent an enormous amount of money on a new campus for the sciences, and this project could be seen as a counterbalance for the humanities.

Or, HYP could band together and put in a mere $100 million each to get the job done.

Is this likely to happen? Of course not. HYP and other wealthy institutions are being asked to spend their prodigious endowments on many other things, and are reluctant to up their spending rate at all. But I believe a HYP or HYP-like solution is much more likely than public

funding for this kind of project, as the Human Genome Project[22] received.

This entry was posted on Thursday, May 29th, 2008 at 3:30 pm and is filed under Books[23], Digitization[24], Google[25], Libraries[26], Microsoft[27], Open Access[28]. You can follow any responses to this entry through the RSS 2.0[29] feed. You can leave a response[30], or trackback[31] from your own site.

References

1. ^ Microsoft (www.microsoft.com)
2. ^ left the business of digitizing millions of books (blogs.msdn.com)
3. ^ they saw it as no business at all (machinist.salon.com)
4. ^ Zotero (www.zotero.org)
5. ^ Internet Archive (www.archive.org)
6. ^ removed the contractual restrictions (www.archive.org)
7. ^ fully open collection (www.archive.org)
8. ^ Open Content Alliance (www.opencontentalliance.org)
9. ^ Universal Digital Library (tera-3.ul.cs.cmu.edu)
10. ^ Google (www.google.com)
11. ^ Book Search (books.google.com)
12. ^ Alfred P. Sloan Foundation (sloan.org)
13. ^ some relationship to all three institutions (www.dancohen.org)
14. ^ MIT's OpenCourseWare program (ocw.mit.edu)
15. ^ Harvard (www.harvard.edu)
16. ^ Widener (hcl.harvard.edu)
17. ^ Princeton (www.princeton.edu)
18. ^ EBay (www.ebay.com)
19. ^ Amazon (www.amazon.com)
20. ^ Yale (www.yale.edu)
21. ^ chose Microsoft (images.library.yale.edu)
22. ^ Human Genome Project (www.ornl.gov)
23. ^ View all posts in Books (www.dancohen.org)

24. ^ View all posts in Digitization (www.dancohen.org)
25. ^ View all posts in Google (www.dancohen.org)
26. ^ View all posts in Libraries (www.dancohen.org)
27. ^ View all posts in Microsoft (www.dancohen.org)
28. ^ View all posts in Open Access (www.dancohen.org)
29. ^ RSS 2.0 (www.dancohen.org)
30. ^ leave a response (www.dancohen.org)
31. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen's Digital Humanities Blog » Blog Archive » Mass Digitization of Books: Exit Microsoft, What Next?*

http://www.dancohen.org/2008/05/29/mass-digitization-of-books-exit-microsoft-what-next/

Readability — An Arc90 Laboratory Experiment

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