Dan Cohen's Digital Humanities Blog » Blog Archive » What Would You Do With A Million Books?

What would you do with a million digital books? That's the intriguing question this month's D-Lib Magazine[1] asked its contributors, as an exercise in understanding what might happen when massive digitization projects from Google, the Open Content Alliance, and others reach their fruition. I was lucky enough to be asked to write one of the responses, "From Babel to Knowledge: Data Mining Large Digital Collections,"[2] in which I discuss in much greater depth the techniques behind some of my web-based research tools. (A bonus for readers of the article: learn about the secret connection between cocktail recipes and search engines.) Most important, many of the contributors make recommendations for owners of any substantial online resource. My three suggestions, summarized here, focus on why openness is important (beyond just "free beer" and "free speech" arguments), the relatively unexplored potential of application programming interfaces (APIs), and the curious implications of information theory.

1. *More emphasis needs to be placed on creating APIs for digital collections.* Readers of this blog have seen this theme in several[3] prior[4] posts[5], so I won't elaborate on it again here, though it's a central theme of the article.

2. *Resources that are free to use in any way, even if they are imperfect, are more valuable than those that are gated or use-restricted, even if those resources are qualitatively better.* The techniques discussed in my article require the combination of dispersed collections and programming tools, which can only happen if each of these services or sources is openly available on the Internet. Why use Wikipedia (as I do in my H-Bot tool[6]), which can be edited—or vandalized—by anyone? Not only can one send out a software agent to scan entire articles on the Wikipedia site (whereas the same spider is turned away by the gated

Encyclopaedia Britannica), one can instruct a program to download the entire Wikipedia and store it on one's server (as we have done at the Center for History and New Media[7]), and then subject that corpus to more advanced manipulations. While flawed, Wikipedia is thus extremely valuable for data-mining purposes. For the same reason, the Open Content Alliance[8] digitization project (involving Yahoo, Microsoft, and the Internet Archive, among others) will likely prove more useful for advanced digital research than Google's far more ambitious library scanning project, which only promises a limited kind of search and retrieval.

3. *Quantity may make up for a lack of quality*. We humanists care about quality; we greatly respect the scholarly editions of texts that grace the well-tended shelves of university research libraries and disdain the simple, threadbare paperback editions that populate the shelves of airport bookstores. The former provides a host of helpful apparatuses, such as a way to check on sources and an index, while the latter merely gives us plain, unembellished text. But the Web has shown what can happen when you aggregate a very large set of merely decent (or even worse) documents. As the size of a collection grows, you can begin to extract information and knowledge from it in ways that are impossible with small collections, even if the quality of individual documents in that giant corpus is relatively poor.

This entry was posted on Friday, March 17th, 2006 at 11:56 am and is filed under APIs[9], Books[10], Google[11], Information Theory[12], Mashups[13], Text Mining[14], Wikis[15], Yahoo[16]. You can follow any responses to this entry through the RSS 2.0[17] feed. You can leave a response[18], or trackback[19] from your own site.

References

1. ^ this month's D-Lib Magazine (www.dlib.org)
2. ^ "From Babel to Knowledge: Data Mining Large Digital Collections," (www.dlib.org)

3. ^ several (www.dancohen.org)
4. ^ prior (www.dancohen.org)
5. ^ posts (www.dancohen.org)
6. ^ my H-Bot tool (chnm.gmu.edu)
7. ^ Center for History and New Media (chnm.gmu.edu)
8. ^ Open Content Alliance (www.dancohen.org)
9. ^ View all posts in APIs (www.dancohen.org)
10. ^ View all posts in Books (www.dancohen.org)
11. ^ View all posts in Google (www.dancohen.org)
12. ^ View all posts in Information Theory (www.dancohen.org)
13. ^ View all posts in Mashups (www.dancohen.org)
14. ^ View all posts in Text Mining (www.dancohen.org)
15. ^ View all posts in Wikis (www.dancohen.org)
16. ^ View all posts in Yahoo (www.dancohen.org)
17. ^ RSS 2.0 (www.dancohen.org)
18. ^ leave a response (www.dancohen.org)
19. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen's Digital Humanities Blog » Blog Archive » What Would You Do With a Million Books?*

http://www.dancohen.org/2006/03/17/what-would-you-do-with-a-million-books/

Readability — An Arc90 Laboratory Experiment

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