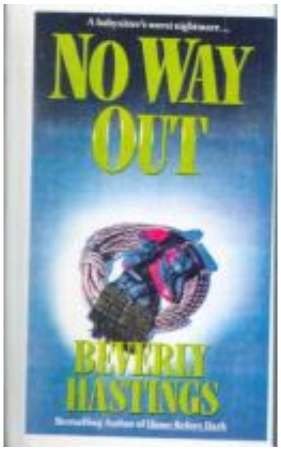
Dan Cohen’s Digital Humanities Blog » Blog Archive » Why Google Books Should Have An API


A babysitter's worst nightmare...
NO WAY OUT
BEVERLY HASTINGS
Bestselling Author of Home Before Dark

[This post is a version of a message I sent to the listserv for CenterNet[1], the consortium of digital humanities centers. Google has expressed interest in helping CenterNet by providing a (limited) corpus of full texts from their Google Books[2] program, but I have been arguing for an API[3] instead. My sense is that this idea has considerable support but that there are also some questions about the utility of an API, including from within Google.]

My argument for an API over an extracted corpus of books begins with a fairly simple observation: how are we to choose a particular dataset for Google to compile for us? I’m a scholar of the Victorian era, so a large corpus from the nineteenth century would be great, but how about those who study the Enlightenment? If we choose novels, what about those (like me) who focus on scientific literature? Moreover, many of us wish to do more expansive horizontal (across genres in a particular age) and vertical (within the same genre but through large spans of time) analyses. How do we accommodate the wishes of everyone who does computational research in the humanities?

Perhaps some of the misunderstanding here is about the kinds of research a humanities scholar might do as opposed to, say, the computational linguist, who might make use of a dataset or corpus (generally a broad and/or normalized one) to assess the nature of (a) language itself, examine frequencies and patterns of words, or address computer science problems such as document classification. Some of these corpora can provide a historian like me with insights as long as the time span involved is long enough and each document includes important metadata such as publication date (e.g., you can trace the rise and fall of certain historical themes using BYU’s Time Magazine corpus[4]).

But there are many other analyses that humanities scholars could undertake with an API, especially one that allowed them to first search for books of possible interest and then to operate on the full texts of that ad hoc corpus. An example from my own research: in my last book[5] I argued that mathematics was “secularized” in the nineteenth century, and part of my evidence was that mathematical treatises, which normally 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 on the writing of high-profile mathematicians.

How would I go about supporting this thesis today using Google Books? I would of course love to have an exhaustive corpus of mathematical treatises. But in my book I also used published books of poems, sermons, and letters about math. In other words, it’s hard to know exactly what to assemble in advance–just treatises would leave out much of the story and evidence.

Ideally, I would like to use an API to find books that matched a complicated set of criteria (it would be even better if I could use regular expressions to find the many variants of religious language and also to find religious language relatively close to mentions of mathematics), and then use get_cache to acquire the full OCRed text of these matching books. From that ad hoc corpus I would want to do some further computational analyses on my own server, such as extracting references to touchstones for the divine vision of mathematics (e.g., Plato’s later works, geometry rather than number theory), and perhaps even do some aggregate analyses (from which works did British mathematicians most often acquire this religious philosophy of mathematics?). I would also want to examine these patterns over time to see if indeed the bond between religion and mathematics declined in the late Victorian era.

This is precisely the model I use for my Syllabus Finder[6]. I first find possible syllabi using an algorithm-based set of searches of Google (via the unfortunately deprecated SOAP Search API[7]) while also querying local Center for History and New Media databases for matches. Since I can then extract the full texts of matching web pages from Google (using the API's cache function), I can do further operations, such as pulling book assignments out of the syllabi (using regular expressions).

It seems to me that a model is already in place at Google for such an API for Google Books: their special university researcher's version of the Search API[8]. That kind of restricted but powerful API program might be ideal because 1) I don't think an API would be useful without the get_OCRed_text function, which (let's face it) liberates information that is currently very hard to get even though Google has recently released a plain text view of (only some of) its books; and 2) many of us want to ping the Google Books API with more than the standard daily hit limit for Google APIs.

[Image credit: the best double-entendre cover I could find on Google Books: No Way Out *by Beverly Hastings.]*

This entry was posted on Tuesday, September 4th, 2007 at 3:20 pm and is filed under APIs[9], Books[10], Google[11], Open Access[12], Text Mining[13]. You can follow any responses to this entry through the RSS 2.0[14] feed. You can leave a response[15], or trackback[16] from your own site.

References

1. ^ the listserv for CenterNet (lists.digitalhumanities.org)
2. ^ Google Books (books.google.com)
3. ^ API (www.dancohen.org)
4. ^ BYU's Time Magazine corpus (corpus.byu.edu)
5. ^ my last book (www.dancohen.org)
6. ^ Syllabus Finder (chnm.gmu.edu)

7. ^ SOAP Search API (code.google.com)
8. ^ their special university researcher's version of the Search API (research.google.com)
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 Open Access (www.dancohen.org)
13. ^ View all posts in Text Mining (www.dancohen.org)
14. ^ RSS 2.0 (www.dancohen.org)
15. ^ leave a response (www.dancohen.org)
16. ^ trackback (www.dancohen.org)

Excerpted from *Dan Cohen's Digital Humanities Blog » Blog Archive » Why Google Books Should Have an API*

http://www.dancohen.org/2007/09/04/why-google-books-should-have-an-api/
