Exhibit P

Daniel Clancy Says:
December 17th, 2008 at 1:26 am

I am Dan Clancy, the engineering manager for Google Book Search. I want to respond to some of the points made in this blog post since some of them are inaccurate and others misleading. Before providing a detailed response, I want to emphasize that Google is very supportive of public debate regarding the various terms of the agreement. However, it is important that such debate is factually accurate. We have offered to meet with the OCA to help clarify questions about the settlement, and we are still hopeful a meeting will happen. More information about the agreement is at http://books.google.com/googlebooks/agreement/ and, if you are an author, publisher or other potential member of the class, you should check out the Class website at http://books.google.com/booksrightsholders/. Now, on to a specific response...

In addition to the benefits described above, our library partners will also receive a copy of the digital files of the books we digitize. The settlement agreement includes explicit authorizations on how they can use these files as well as explicit restrictions. In some cases, these explicit authorizations will result in increased usage due to the certainty that they offer to our partners. While there may have been some uses that our library partners may have considered under our old agreement that are prohibited now, in most cases the need for these uses is replaced by the subscription service that Google is offering them at no expense. Furthermore, as noted above, many of the uses offered by the subscription service would not have been possible in the old regime due to copyright law. In addition, the scope of the copies that are returned to the libraries is greater. Under our current agreement, a library partner only receives copies of files from books we scan from that library. In the settlement agreement, for partners that go over certain thresholds of scanning, they will be able to receive copies of books that we scan from other libraries. This ensures that multiple parties will have copies of these files to preserve for posterity. As always, it is important to note that our library partnerships are non-exclusive. This agreement does not restrict a library in any way with respect to other digitization initiatives that they may consider.

Open Content Alliance (OCA) » Blog Archive » A Raw Deal for Libraries
http://www.opencontentalliance.org/2008/12/06/a-raw-deal-for-libraries/#comment-232
Screen clipping taken: 9/1/2009, 5:21 PM

Open Content Alliance (OCA) » Blog Archive » A Raw Deal for Libraries
http://www.opencontentalliance.org/2008/12/06/a-raw-deal-for-libraries/#comment-232
Screen clipping taken: 9/1/2009, 5:20 PM

Exhibit Q

Language Log

- Home
- About
- Comments policy

The Google Books Settlement

August 28, 2009 @ 10:29 am · Filed by Mark Liberman under Books

« previous post | next post »

I'm spending today at Berkeley, participating in a one-day conference on "The Google Books Settlement and the Future of Information Access". I'll live-blog the discussion as the day unfolds, leaving comments off until it's over. I believe that the sessions are being recorded, and the recordings will be available on the web at some time in the near future. [Gary Price at Resource Shelf provides some other links here, and a press round-up here. Another summary by an attendee is here.]

Regular LL readers will know that we've been long-time users and supporters of Google Books, with occasional complaints about the poor quality of its metadata. For a lucid discussion of some issues with the terms of the proposed settlement, read Pamela Samuelson's articles "The Audacity of the Google Books Settlement", Huffington Post, 8/10/2009, and "Why is the Antitrust Division Investigating the Google Books Search Settlement?", Huffington Post, 8/19/2009.

[Note that in most cases below, first-person pronouns refer to the speaker, not to me.]

The opening panel is on the topic "Datamining and non-consumptive use", and the panelists are Peter Brantley and Jim Pittman, with Eric Kansa as the moderator.

AnnaLee Saxenian, Dean of Berkeley's School of Information, kicked things off by noting that this all started five years ago at the Frankfort Book Fair, where Google announced what was then called "Google Print". In 2005, the Authors Guild and a number of publishers separately sued Google. A proposed settlement will be considered for approval in New York district court in October. The purpose of this conference is to have a conference about the opportunities and risks of this settlement.

Eric Kansa — Google has publicly said that 7 million books had been scanned as of some time in 2008; some others estimate up to 15 million now. This is a large piece of the total universe of book — US Census reports 2.334 M books published in the U.S. between 1880-1998, WordCat lists 23 million books.

Comparison and contrast to Human Genome project, which had significant public sponsorship to create; the result is largely held in public trust. Google Books is being created with private funds and will largely be privately held, with restrictions imposed by the settlement.

The largest value of this corpus might lie in the results of machine processing, even if no human eyeballs ever viewed it. These "non-display" or "non-consumptive" uses include arbitrary uses allowed within Google. For others, a "research corpus" will be created and housed at two US-based participating libraries. Qualified researchers can experiment with the data for non-commercial purposes. The two sites will be responsible for evaluating research proposals, running audits, etc.

Traditionally, US law makes a distinction between public-domain "facts" and "ideas", and copyrighted expression of such facts and ideas. The settlement appears to give rights over information extracted from the corpus to Google, the Authors Guild and the publishers.

Peter Brantley — representing the "Open Book Alliance" as well as the Internet Archive. I.A. Richards: "A book is a machine to think with". That takes on a new meaning, because we're now moving from "books as books" to "books as data". Future books will be different, in ways that most publishers and lawyers don't seem to be thinking about, and in fact no one really can predict. Moving from Gemeinschaft to Gesellschaft.

Concerns: Corpus is unique — a comprehensive collection of 20th-century literature, based on an exclusive deal, due to the treatment of "orphan works" as well as other aspects of the legal settlement. Do we want to give sole ownership of this unique resource to a single corporate actor? Even if authors pull their books from the display portion, it will still be there in the datamining portion. Prices charged to universities for this unique resource will be unregulated and unrestrained by competition.

Other scenarios: should there be some sort of compulsory licensing? Choice shouldn't just be settlement/no settlement. It's a historic moment. How can we protect the future?

Jim Pitman — The raw content of the books is a tremendous resource. But there's also the information that results from the interaction of users with this content, or from algorithmic analysis of the content, or both. JP is especially interested in subject navigation and classification. Would development of a "map of knowledge" form this corpus belong to Google? To what extent will the scholarly community be inhibited in developing such ideas by the terms of this settlement?

Do we have to ask for permission to pull out (say) a tableau of theorems, or a table of compounds?

There's lots to be done by text analysis, but you can't really do it right if you can't get access to the data on your own computer. Going to one of two designated centers to run some code in a limited time period on someone else's machine is hardly even second best.

Google has an intrinsic commercial interest in stifling innovation by others in textual analysis — at what point will they give in to the temptation to act on this interest?

To get the best results, you want to combine machine learning with human judgment. To what extent will the terms of the settlement allow this? Individuals and small organizations should be encouraged to cultivate the parts of the academic universe that they know best — Google books will be a big part of the resources for this, but it should be easy to combine it with other things. Does the proposed settlement inhibit this?

Dan Clancy of Google Books responds: Nonconsumptive access is fair use, in Google's view. So you don't have to ask permission of Google to do whatever [but you have to have the texts — myl]. Control of research access will be up to the two host sites. There are some provisions, when it talks about the information extracted, that we still don't understand as a community. But the idea is that (for example) if you use the corpus to infer a subject classification system — that's yours. But if you want to use the corpus to derive an index, and offer access to books on that basis, you can't. Similarly, you can't offer a concordance to (say) works belonging to Elsevier.

Jim Pitman: what about an indexing service for mathematical theorems?

Dan Clancy: theorems don't belong to us; but if it's a commercial service, we don't have the rights to give you the right to link to someone else's text. And if we aren't already offering a theorem search service, it's not competing, and we wouldn't be able to object.

Jim Pitman: Is there a list of Google services that we can't compete with?

Dan Clancy: No.

Jim Pitman: What about an algorithm to disambiguate references to people, to be used in other data mining applications?

Dan Clancy: I don't think that would be any problem.

Eric Kansa: What about authors' or publishers' rights to withdraw data-mined facts from downstream applications? And do you need permission to use extracted information in a commercial application?

Dan Clancy: The license in this case adds additional constraints, beyond copyright law, so that yes, the corpus can only be used for research, and if you want to use the results of algorithmic processing of the corpus in a business, you'd need to get permission from the Authors Guild and perhaps others.

Jason Schultz, a clinical professor at the Berkeley Law School: What's interesting about the settlement is its breadth — it sets up a legal regime by which Google is bound, and by which lots of others outside the lawsuits are bound. Things could get changed and a second version of the settlement could come around.

What happens if the settlement is rejected? Probably Google book search goes forward as it is now.

Dan Clancy: But in that case, there's no research corpus.

Jason: If the settlement is rejected and Google wins the case, that would be a precedent that others could rely on. Also, state universities might have sovereign immunity.

Marty: Does this provide a solid enough basis for ongoing research, including the ability to reproduce others' research? What sort of access will researchers really get, and to what? This makes a big difference to the kinds of research that can be done?

Dan Clancy: The simple answer is you get full access. But the question of how to give access to researchers to compute at scale over the research corpus is unclear. We see the two access sites as an open source community.

Jamie Love: it would be helpful to elaborate on what research institutions would NOT have the freedom to do?

Dan Clancy: No restrictions for noncommercial services; except it can't be used to compete with Google.

The second panel is on privacy issues — that is, who gets to know what books someone reads.

The first speaker is Angela Maycock: The settlement is silent on the issue of privacy. There have been lots of informal assurances, but that's all we have to go on.

Why do librarians care? What is the history? What are the threats? How does a loss of privacy rights affect users? What are the implications for this discussions?

In a library context, "Privacy" is the right to engage in open inquiry, without having your activities scrutinized by anyone else. "Confidentiality" is the right of libraries to protect users' information. Lack of privacy and confidentiality has a chilling effect, which damages first-amendment rights.

Examples of challenges — needed to counter the misconception that "good people don't have anything to

hide". In Wise County TX, the DA asked the library for all people who checked out a book on childbirth during the previous nine months, as part of an investigation of a child abandonment case. In Colorado, the supreme court ruled that the DA could not access someone's history of purchases from the Tattered Cover bookstore.

The American Library Association has not opposed the settlement, but did file a letter asked for vigorous oversight. Google people have been good listeners.

Tom Leonard: Research libraries are generally not exposed to the kinds of pressures that public libraries are. Still, the core objective for research libraries remains preventing users from being monitored, except in exceptional cases where the user is fully informed.

Readers do know that we've kept a record of what they've borrowed, that this information will never be shared with anyone, and that it will be purged after the book is returned. Similarly, IP addresses are purged as soon as possible after use of online databases. The only exception is the case of rare and valuable books, were records are kept in order to guard against theft.

There are some hard cases, e.g. suppose a group abroad is using our resources to assign people to "untouchable" categories, by looking at census data and maps and so on, should that fact be conveyed to those who are affected?

What about after the GBS? We have the capacity now to monitor all digital searches, but we don't do it, and we shouldn't do it. I believe Google when they say that they feel the same way, but I would like to believe them more.

Jason Schultz: Stanford marshmallow experiment — We have a settlement in front of us: how good will this marshmallow be? How much better (or not) would things be if we wait?

When GB first came on the scene, the issues were all about copyright and fair use. I'm a big fan of the fair use argument for scanning books for information access, information location. This is the right balance in copyright law, and an excellent way to address the orphan works problem, at least in part.

But now we're in a different situation. The settlement deals with so much more than just copyright and fair use. Four things relevant to privacy:

1) The size of the deal. This is the largest copyright licensing deal in history.
2) The compulsory nature of the settlement — it includes everyone who has a copyright interest in the U.S., and it's binding on everyone unless they opt out: "You are giving Google permission, and in return, you get some benefits". Now Google offers GMail, and you can take it or leave it. But here they need permission from more people than have ever been involved in such a process before. Is privacy important enough to be part of the deal?
3) Implications for privacy. If Google can look at every page that you ever read, ??
4) This is a legal hack — which I like in general, EFF often used them — but this is a VERY big hack.

It's not like the privacy issues with the Amazon Kindle or the iPhone, because it's no much bigger, so much longer term, and it's compulsory.

Are the enforcement and accountability mechanisms adequate in this case?

Two models that might be followed:

All of the records in Google Health are kept separate from all your other Google info.

Google's location product — Latitude — has amnesia built into it, it only knows where you are now, not where you were.

Should Google Books do similarly?

If privacy violation occur, how will we know? What will the consequences be?

Michael Zimmer: Planet Google becomes the center of gravity of everyone's information-seeking activities; results are good, which is why we do it, but the resulting infrastructure is a potential threat. This is why we're worried about data retention policies, etc.

Norms of information flow: when you go into a library, in real life or on line, you have some expectations about privacy. Norms for web searching are different. A certain amount of tracking is inherent — and beneficial. A lot of people accept it and move on. Now the question arises, which norms apply to looking at Google Books on line?

Competition: when users are informed, they can "vote with their feet" if they don't like the privacy policy of a given provider. But Google Books will be a monopoly — if people don't like it, where can they go?

There's an FAQ on the Inside Google Books blog. A Google Account will not be necessary; they won't sell access information; their general privacy policy applies.

I trust Google; but like Tom says, I'd like to trust more. Google Street view had very different policies in the U.S. vs. in Europe and Canada, because privacy laws were different. But maybe the ethics should be the same in both cases.

Options: Build in anonymity online? Should book search data be protected the way that health data is? Should everything?

The first afternoon session is on Quality.

Paul Duguid is the moderator.

The panelists are me, Geoff Nunberg, Cliff Lynch, and Dan Clancy.

My presentation is here.

Summary of Geoff Nunberg's talk: This is "the last library". So it's important to get it right. But dates, authorship information, categories are often pretty bad.
[Details are in Geoff's slides here.]

Cliff Lynch: It feels like something is happening here that is much bigger than a simple legal settlement. We're making a national decision about what may well be, as Geoff said, "the last library" — and it's important to get that right.

Metadata problems are much less excusable than scan and OCR problems. OCR is a an error-prone process that gets better as algorithms improve. But bibliographic metadata is a mature and well-understood area — if this is "the last library", why not take the trouble to get it right? Three kinds of things mixed up together — OCR output, assertions from bibliographic metadata, assertions from publisher metadata.

Dan Clancy: Some of the most vocal critics are also some of the most avid users. It's important to understand that dialogue and criticism is an important thing.

Is this about identifying problems and thinking through solutions? or is it about creating a dichotomy? How do we move forward?

I actually don't view Google Books as the one and only library. I don't think it will be and I don't think it should be. There will continue to be other digitization activities. To the extent that Google Book Search is our last shot, then without Google Book Search we never would have had a shot; and I don't really think that's true.

For books scanned at a library, we get the metadata from the library. We get updates from libraries every week or two to fix publication date issues. We probably do have a problem with classification, though it's not done automatically. We get classification metadata from many sources, and perhaps we're not combining them in the best way.

To the extent that the source of our data is part of the problem — which it mostly is — now we need to think about how to make things better. Merging records is a tough problem. But part of this is not about Google doing this, it's something that our partners need to do as well.

Metadata is only the tip of the iceberg — there's also public-domain determination. If we trusted our metadata about what was before 1923, our error rate would be atrocious, so we can't afford to do that.

The Human Genome project is not really comparable, because there's no competition in this case, just cooperation. We have partnerships with Hathi Trust, with Michigan, etc. It's not a competition, where as one gains the other loses. It's really about figuring out how you can build a bunch of repositories, and how do you ensure from a preservation perspective that books are held in multiple locations.

I don't think we're the only library, we're part of a broader community.

[myl: but Dan said later "Do you think that if Google hadn't organized this scanning effort, someone else would have done it? Do you think that if Google hadn't done it, someone else would do it in the next 10 or 20 years? I think the answer to both questions is 'no'".]

Ed Feigenbaum: tens of millions of dollars is a relatively small amount of money. Why is this the last library?

Someone else: Who owns the metadata? Who controls it? Under what conditions can it be distributed? Who can change it and re-distribute the changed version? Is it OCLC? Is it the contributing libraries? What about Google's contribution to merging? Cliff Lynch: it's a mess. Nobody really knows.

Ed Clancy: It's more than tens of millions of dollars, alas.

Guy from EFF: We've suggested that Google should escrow these scans, and after a reasonable time (say 28 years, which was good enough for the copyright act for our founding fathers), make them available to others.

Guy from CNET: How much has Google spent?

Dan Clancy: Alas, we do not publicly disclose the amount of money. We've scanned about 10 million books. Internet archive's cost is said to be about $30 per book.

The last panel is on public access, moderated by Pamela Samuelson.

The first speaker is Dan Greenstein: Why did UC libraries get involved? Libraries are all about access to information, and public libraries including libraries in public universities see public access as a sacred trust. UC libraries were active in the settlement, because of the benefits to the public.

The second speaker is Carla Hesse. Her remarks focused on the question of whether Google Books might turn out to be "too big to fail", and to require a government bailout at some point in the future, perhaps because that slice of Google's business might be sold off to some third party who is less successful at deriving revenue from it, or perhaps for other reasons.

The third speaker is Jamie Love, who offered a detailed and interesting critique of the economics of the proposed settlement. I'll try to get his slides, since there were a lot of details that are hard to get down in real time. One key point is that the settlement allows publishers to collude to fix prices, in a way that would be clearly a violation of anti-trust if they did it on their own, outside of the context of the settlement.

The fourth speaker is Molly Van Houweling. Her remarks focused on the question of what future pricing and subscription options for university libraries might look like.

Dan Clancy: Carla did a good job of broadening the scope — she placed this in the context of the broader evolution in how we access information. Now we're not dealing with physical goods, but rather with digital goods; the book search settlement is not mainly about that, it's mainly about orphaned works and so on. Future books will mainly be distributed in digital form by publishers. We don't see this settlement as the overall solution, just as the solution to the specific problem of out-of-print books that are still in copyright.

Responding to Jamie Love, many months of the settlement discussion were devoted to the questions of competition and pricing. Most of what people want and where the competition is, is for in print books. What happens in the journal market is you have publishers who don't even let the authors distribute their works. But in the settlement, all authors are permitted CC distribution. The only place you can get the latest and greatest is from the publisher. For Google Books, there are lots of places you'll be able to get access to most of these books. For example from the library as a physical copy. For digital copies, the internet drives prices down, and our users will get free search, free preview, and low purchase prices (mostly $14 or less, whereas interlibrary loan is mostly $20 or $30).

Questions from audience:

What about access to Google Books from outside the U.S.?

Dan Clancy: U.S. Courts can only authorize use within the U.S. But as rights-holders are identified, they can choose to authorize access outside the U.S.

What about the role of hackers and pirates? In music and movies, the best metadata is created by hackers and pirates? When Google's DRM is broken, what will happen?

Dan Clancy: DRM is not worth much, true, but the long tail means that for most of the books of interest, you won't find them on a local peer.

Pam Samuelson: There are complicated, elaborate, strong, expensive-looking requirements for data security on the part of the universities hosting the "research corpus" infrastructure. This could get to the point where the research host sites will have to close down for lack of funding.

Because I'm a lawyer, I look at "what could go wrong?" There are lots of places where things could break, and the security responsibilities that the host universities have to undertake is a potential failure point.

Dan Clancy: Those standards were written so as to consistent with current library practices. Libraries wanted to move from a statutory damages regime to an actual damages regime, which the settlement accomplishes. The likelihood that the actual damages will be significant is very very small.

[Mark Liberman: There were a number of additional interesting points in the discussion that I didn't get in time — if you're deeply interested in the topic, you'll find them in the recording that will be on line in a week or so. I'll add a link here, as well as blogging it separately when it becomes available..]

August 28, 2009 @ 10:29 am · Filed by Mark Liberman under Books

Permalink

---

6 Comments »

1. **John Cowan said,**

   August 29, 2009 @ 1:01 am

   Sounds like everybody's thinking and talking, as opposed to retreating to prepared positions and bashing each other over the head, as in the rest of what passes for debate in the world today. That's very refreshing.

   (Disclaimer: I work for Google, but have only the tiniest of connections with Book Search, and don't know anything about it that is material nonpublic information.)

2. **Jay Lake: [links] Link salad for a quiet Saturday morning said,**

   August 29, 2009 @ 10:17 am

   [...] Language Log with lots of neepery on the Google Books settlement — Regular readers are aware that I consider the Google Books settlement to be nothing more than rancid, institutionalized thievery, in direct contradiction to the company's famous "Don't be evil" rubric. [...]

3. **ResourceShelf » Blog Archive » Live From Berkeley: Google Books Settlement and the Future of Information Access Conference said,**

   August 29, 2009 @ 10:43 pm

   [...] Mark Liberman at Language Log is doing an excellent job live blogging the event. [...]

4. **Tadeusz said,**

   August 30, 2009 @ 6:46 am

   Many thanks for the excellent coverage, but I have a linguistic question: as a non-native speaker of English I was a bit puzzled by the first sentence:
   "I'm spending today at Berkeley".

   I did find two occurrences at http://www.americancorpus.org/:

   our way back to Haleiwa, I'm glad all over again that I'm spending
   today with Kai. When I was younger, I thought birthdays were so

special
(this is pretentious a bit)
California's a long way. # So Al Scott's spending today helping a
13-year-old named Jason McGuire who attends Brown Junior High.
(this is ambiguous, actually)

I take it that in the sentence "today" functions as a noun, and "today" in the nominal function occurs above all in set expressions: "of today", "today's".
I wonder to what extent the construction "I am spending today" is standard (normal, whatever the adjective)?
Thank you.

5. **Avi Rappoport said,**

August 30, 2009 @ 12:56 pm

Thanks for the notes. There are so many issues here it's hard to get my head around them. The monopoly problem is the one that still worries me, with too many economic incentives for both the Author's Guild and Google to abuse their position.

PS, Tadeusz, on a practical level "spending today" is the same as "spending the day".

6. **Digging Digitally » Google Book Settlement Follow Up said,**

September 1, 2009 @ 1:54 pm

[...] I've had a chance to digest our recent conference on the Google Books Settlement. Like many other observers, I came away from the event less clear about what the Settlement actually means and how it will shape the future landscape of information access. Mark Liberman, a conference participant and pioneer in computational humanities (and other areas) live-blogged the event here. [...]

RSS feed for comments on this post · TrackBack URI

Leave a Comment

Name (required)

E-mail (required, never displayed)

URI

Submit Comment

- **Search**

Search

Archives [+/–]

-
- [Posts before 4/8/2008 are here]
  [Search old posts here]

List of authors:

-
  - Barbara Partee
  - Benjamin Zimmer
  - Bill Poser
  - Chris Potts
  - David Beaver
  - Eric Baković
  - Geoff Nunberg
  - Geoffrey K. Pullum
  - Heidi Harley
  - John McWhorter
  - Mark Liberman
  - Melvyn Quince
  - Paul Kay
  - Roger Shuy
  - Sally Thomason
  - Steven Bird
  - Victor Mair
  - Zwicky Arnold

Other authors [+/–]

-

Blogroll [+/–]

-

Categories [+/–]

-

- **Meta**
  - Log in
  - RSS 2.0
  - Atom
  - WordPress

Powered By WordPress