& Library Digitization

METRO "Google and Libraries"
Columbia University, March 2008

Laura Quilter*
lquilter @ lquilter.net
http://lquilter.net/talks/2008-glib.html

* unpublished contract analysis data from Jack Lerner & Laura Quilter
Roadmap

- Overview of digitization projects
- Impact, critiques, responses, competitors
  - Impact on digitization initiative
  - Privacy concerns
  - Propertization of information concerns
  - Cultural behemoth concerns
  - Lawsuits
- Contracts close-up
Google’s Book Search

- Library digitization program announced December 2004
- 5 libraries initially: UMich & Stanford (all), Harvard (particular collections), NYPL & Oxford (public domain); scores of libraries since, including as of December, Columbia
- # of books so far: Initially planned 15 million; latest estimates suggest 32M in 6 years. UMich recently hit first million.
Ripple of Effects

- Massive press coverage (of course)
- Publishers and Author’s Guild filed lawsuits
- New digitizing initiatives:
  - Open Content Alliance, Microsoft, European, China/Library of Alexandria, Carnegie Mellon’s “Universal Library”, and many others
  - Related: Google Scholar
- Google’s Book Search goes online
- “Unauthorized”: LibraryThing, Ultrapedia
- Scholarly ink begins spilling in earnest
What does GBS look like?

- Google full-text searching
- List of results
- Full-text, licensed text, “snippets”
- List view or cover view
- Highlighted PDFs or OCR plain text
- Resolution - not archival quality; “quantity over quality”
  - see Paul Duguid, *First Monday*, v.12, n.8 (2007), examining Sterne’s *Tristam Shandy*
- Features: LCSH, “My Library”, etc.
What does GBS look like?
... not exactly archival quality

Emma Goldman, *The Social Significance of the Modern Drama* (1910), pp. 274-75
... not exactly archival quality

The Gentleman’s Magazine, 1855 (Siva’s Googlization of Everything blog, 2007/12/5); see similar error at an 1888 edition of Plot’s The Trial and Death of Socrates.
“Plain” text is truly plain

... and dependent on scan quality

Emma Goldman, *The Social Significance of the Modern Drama* (1910), pp. 274-75
(report problems)

Emma Goldman, *The Social Significance of the Modern Drama* (1910), pp. 274-75
Comparisons to other projects

- Library-initiated projects, including Open Content Alliance ("OCA"), use a higher-resolution, "archival quality", and more expensive process.
- No comprehensive comparative evaluation of scan and image qualities as of yet.
- Competitor projects (OCA) "opt-in".
- Public domain only (OCA, MS) (e.g., Brit. Lib.).
- Difficult to compare numbers with consortia and branches, but dozens, maybe scores.
What about the library’s copy?

- Depends on contract
- Most contracts provide copy back to library
  - CIC copies “escrow”
  - PDF + OCR, but no “map” linking index to PDF page location
- Significant contract restrictions (later)
Critiques: Privatization

- Siva most eloquent exponent of this critique
- A task ideally by and for libraries & public
- Where is the library mission?
- Non-exclusivity raised, but multiple mass-digitizations impractical:
  - Expensive: Libraries bear costs of digitization
  - Preservation: Wear & tear on materials
  - Inertia: Sunk costs but unwilling to “re-do”
Critiques: Privacy

- Ties to user accounts: Google ID login required for some copyrighted works
- Protections in libraries of user viewing information: ALA, state statutes, constitutional law -
  - *Neither applicable to, nor followed by, Google*
- Privacy, Speech, Access
  - Accountability vs. restrictions vs. anonymity
  - Fine grained knowledge of consumption
    - What does it lead to?
Critiques: Etc.

- Lack of openness: contract terms, scanning information and quality, list of books
  - “as open as it wants to be”—e.g., LibraryThing
- “Sucking the air” from other digitization projects
  - Cultural imperialism
- Poor-quality scans and difficult works problems
  - e.g., multi-volume works
LibraryThing

Monday, September 17, 2007

Google Book Search ... on LibraryThing

Introducing something new we're calling "Google Book Search Search."

Google Book Search Search is a bookmarklet that searches Google LibraryThing Google Book out for coffee

Tuesday, September 18, 2007

Google Book Search Search...

I am voluntarily and temporarily suspending Google Book Search Search, our effort to distribute the task of collecting Google books IDs for LibraryThing members through a browser "bookmarklet."

I am talking with Google about some linking to Book Search pages. Google LibraryThing and libraries to link to.

Thursday, September 20, 2007

Link LibraryThing accounts to Google?

As I said in my talk post, we have spoken to Google about how to link and search Google Book Search reliably and effectively from LibraryThing.

Unfortunately, I am not at liberty to discuss much more than that. I can say that there is no substance to the rumor that Google is re-engineering CueCats to beam targeted advertisements onto your bedroom wall. I am also able to concede that the press accurately reported how Larry and Sergey beat me at drunken thumb-wrestling. But I cannot comment on whether Abby, sober and wielding a hitherto-unnoticed sixth finger, restored LibraryThing's honor.
Impact on digitization initiatives

- Significantly jump-started the “market”
- Numerous ancillary or competitive projects
  - compensate for gaps (e.g., archival quality; other mtrls)
  - take advantage of perceived vulnerabilities (© - MS)
- Likely impact on related efforts: e.g., Kindle
  - Economic research on galvanic / first-starter effect

- Also to some indeterminable extent simply zeitgeist: Full-text market significantly delayed.

Digitization projects  |  Responses, Impacts, Critiques
Critique: Copyright

● Google’s copy:
  ○ Authors & Publishers: If there is a profit to be made from indexing, we have a right to it.
  ○ On the flip-side, the Internet. This type of full-text indexing is precisely what makes Internet search engines viable; Google extending into material space relying on fair use.
    ■ Siva has cogently described this as “betting the Internet”

● Library copy: ???
Google’s accommodations to copyright

● “snippets”
● Opt-out - copyright-holders can ask for takedown (worked in *Blake A. Field v. Google*)
● *Library contractual restrictions* - Serve a variety of functions, including competition, but also some copyright issues.
● Copyright notices
● Restrictions on downloads
● Temporary moratorium on some works
Major copyright litigations

- Two lawsuits, Assn of American Publishers and Authors Guild, filed in SDNY soon after announcement; consolidated in 2006.
  - Publishers continue to cut deals with Google
  - Broad class definition by authors, citing all copyright-holders at UM, was roundly criticized; see, e.g., Lawrence Solum
- German case (WBG pubr) dropped in 2006; French case & English rumblings
Legal claims

- SDNY: Second Circuit
- Reproduction; Distribution; Display; (security)
- Seeking permanent injunction
  - “from, in any manner, reproducing, publicly distributing and/or publicly displaying all or any part of any Publisher’s copyrighted works….except upon the express prior authorization…”
  - Destruction of all copies held by Google
Fair use defense

- Purpose and Character
  - Transformative; Commercial
- Nature of the Work
  - factual > creative
- Amount and Substantiality of Portion
- Market effect
  - “traditional, reasonable or likely to develop”

See Jonathan Band’s analysis for ALA
Precedents

- **Kelly v. Arriba** (9th Cir)
  - Thumbnails
- **Sega v. Accolade** (9th)
  - Intermediate copying
- **Perfect 10 v. Google** (9th 2003)
  - Thumbnails
- **Field v. Google** (DNV)
  - Opt-out

- **MP3.com** (SDNY)
  - “Retrans[mission] in another medium” not protected
  - Entire CDs
  - Usurping potential future market
  - Distinguishing software transitory copying cases

- **American Geophysical Union** (2nd Cir)
  - Circulating entire copies in for-profit co.
- **Boces** (WDNY)
  - Loaning copies of broadcast recordings

Digitization projects | Responses, Impacts, Critiques
Major copyright questions

- Scope of intermediate copying
- Scope of indexing
- Paper v. digital copies
- FU and market exploitation boundary
  - Market in snippets?

New Yorker, "Google’s Moon Shot"
Critiques: Contracts

- Often not publicly accessible
  - Publicly funded institutions available thru FOIA/state sunshine act laws
- Substantive terms of concern: various restrictions on use
- Non-exclusive but a significant cost to “do over”
Contract terms

- Restrictions prohibiting libraries from making copies publicly downloadable
- Restrictions prohibiting libraries from making their own digital copies fully searchable or indexable
- Limitations on noncommercial use
- Restrictions limiting libraries’ right to lend or share their digital copies with other institutions
Contract term: Public downloads

- **Language:**
  - “University shall develop methods and systems for ensuring that substantial portions of the University Digital Copy are not downloaded from the services offered on University's website or otherwise disseminated to the public at large.”

- **Practical Impact:**
  - Non-user and sometimes user downloads directed to Google, even for public domain works
  - Limits other services, such as print-on-demand, even for public domain works
Contract term: No indexing

● **Contract:**
  ○ Google does not provide the image map coordinates to libraries in many of its contracts. UC was a recent exception.
  ○ Partner libraries may not contract with 3rd parties to index their copies.

● **Practical Impact:**
  ○ Diverts users to Google to provide more complete search results, e.g., highlighted text in PDF
  ○ Libraries may not make their copies broadly available; they are excluded from search engines such as crawlers and indexers. This maintains Google’s search engine monopoly over the copy. *Compare OCA - multiple access points; metadata publicly open for indexing.*
  ○ Deep linking to support online research won’t work without image map coordinates
  ○ Library assessment of its search and use data is significantly hampered
  ○ Development of new library-initiated tools are significantly hampered
Contract term: Limitations on Commercial Use

- Regardless of public domain status of the work, no commercial use of the digitized materials by the library.
- Libraries may not therefore lend or share copies with for-profit libraries, and prevents access by other search engines, and various library services businesses that exist now or possibly in the future.
- Compare: Microsoft contract with CDL similarly forecloses commercial use but permits commercial licensing that benefits Microsoft--so long as it doesn’t benefit Microsoft’s competitors.
Contract term: ILL Restrictions

- Regardless of public domain status of the work, libraries are sharply restricted in their ability to lend copies of digitized works to other libraries.
- Limits on how much can be loaned or shared; e.g., UC < 10%
- Libraries must ensure that other restrictions (downloading, indexing, distribution, etc.) are extended to receiving institutions
- Three available contracts (CA, TX, WI) require separate agreement between Google and any receiving library
- Practical effect:
  - Late-comer libraries who have a copy in their collection, that was previously digitized by Google, will not have access to that copy.
  - Libraries that received poor-quality scans from Google may not swap out and get a better-quality scan from a later library
Jeff Ubois & Peter Kaufman guidelines:

Libraries should be prepared to “negotiate intensively” regarding:

- Limited confidentiality: Partner’s policy regarding information about the partnership
- Complete deliverables -- specified quality and formats and associated data
- More open access for library’s users
- Rights to cooperate with other institutions -- commercial & noncommercial
- Responsible treatment of usage data
- Limited duration and survivability: Restrictions should terminate when the agreement does.
- How flexible the deal is to adjust, improve, or abandon
Contract negotiations

Richard Johnson / ARL:

Principles

- Digital libraries are libraries
- Digital content must be treated like other library materials regarding ease of access, freedom of information, preservation, collection development
- Digital activities & collections sustainability -- funding, technology, etc.
- Ongoing communication needs.
- Potential international audience
- Educated workforce & continued learning to support them
- Preservation is critical; development of standards and funding models
- Standards for digital collections
Contract negotiations

- Negotiation checklist:
  - Exclusivity
  - Uses of the digital files
  - Respect for the public domain and copyright
  - Preservation
  - Standards
  - Quid Pro Quo
  - Transparency