Massively Digitizing UC Collections

Ivy Anderson
Director, Collections
California Digital Library
May 2009

Outline

- History and overview of current projects
- What we have digitized and where you can find it
- Use cases
- Google Settlement overview and implications

Brief History of Library Digitization

Berkeley, University of California, Bancroft Library, UCB 150, f. 252v

- Special Collections
  - Manuscripts, archival collections, photographs, etc.
- CDL / UC Libraries
  - Online Archive of California
  - Calisphere

Brief History of Library Digitization

- Specialized Texts and Corpora
  - Making of America -10,000 texts in 10 years
- CDL
  - eScholarship Editions

Brief History of Library Digitization

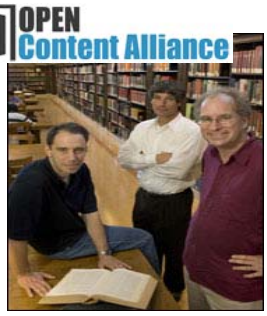
Satans stratagems, 1648. copy from UCLA Library

- Commercial Partnerships
  - EEBO: 100,000 important early English texts
  - Licensed access via ProQuest
    - (restricted)

…and along came Google

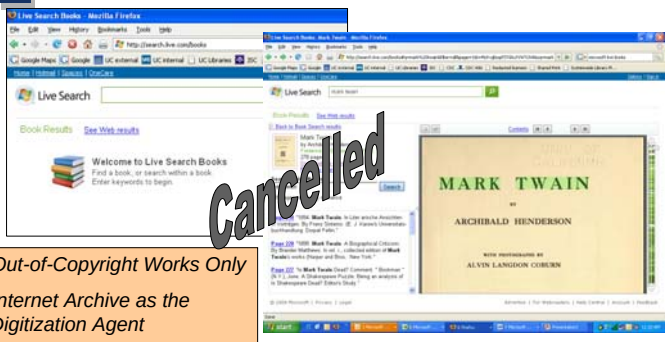
- Google Library Project
  - *2005:* The 'Google Five:'
    - *Harvard, Oxford, New York Public Library, Stanford, University of Michigan*
  - *2006: UC joins*
  - *2009:* 22 library partners in 5 countries
  - 7 million volumes in 4 years
- Google Publisher Partner Program

…and the Open Content Alliance



- *October 2005*
  - Founders: Internet Archive, University of California, U of Toronto…
  - Large-scale digitization of *out-of-copyright works only*
  - A project of the Internet Archive

…and for a time, Microsoft



- *Out-of-Copyright Works Only*
- *Internet Archive as the Digitization Agent*
- *Project Cancelled in 2008*

UC Mass Digitization Timeline

Founding Member of Open Content Alliance	UC Joins Google Library Project	Microsoft Digitization Agreement
October 2005-	August 2006-	March 2007 -July 2008

Project Characteristics

- Google
  - 1.8 million books digitized to date
  - In-copyright and out-of-copyright works
  - All languages
  - Foldouts are skipped
  - Scanning done at a Google off-site facility
  - Funded entirely by Google
- Internet Archive / Open Content Alliance
  - 200,000 books digitized to date
  - Out-of-copyright works only
  - Primarily English language, some romance languages now
  - Foldouts are included
  - Scanning done on-site at SRLF
    - Formerly also at NRLF
  - Previously funded by Microsoft, Yahoo, Sloan Foundation
  - **Future funding uncertain - likely to require library and/or grant funding**

Participating UC Campuses

- Google
  - Northern Regional Library Facility (NRLF)
  - UC Santa Cruz
  - UC San Diego
  - Planned: Bancroft, UCLA, additional campuses
- Internet Archive
  - Northern Regional Library Facility (NRLF)
  - Southern Regional Library Facility (SRLF)
  - UC Berkeley, Bancroft Library
  - UCLA
  - UC Davis

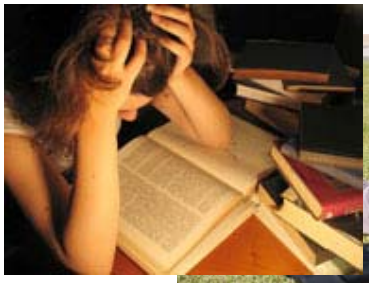
What material has been digitized to date?

- 2 million UC volumes digitized as of April 2009
- Google
  - NRLF: all subjects
  - UC Santa Cruz: Humanities / Social Sciences
  - UC San Diego:
    - East Asian collection
    - International Relations Pacific & Studies
    - Scripps Institute of Oceanography
- Internet Archive / OCA
  - Pre-1923 English language books from NRLF and SRLF
  - Pre-1923 foreign language books at SRLF (in process)
  - UC Davis: Selected California documents
  - Selected Bancroft and UCLA collections
- More UC libraries and collections are planned

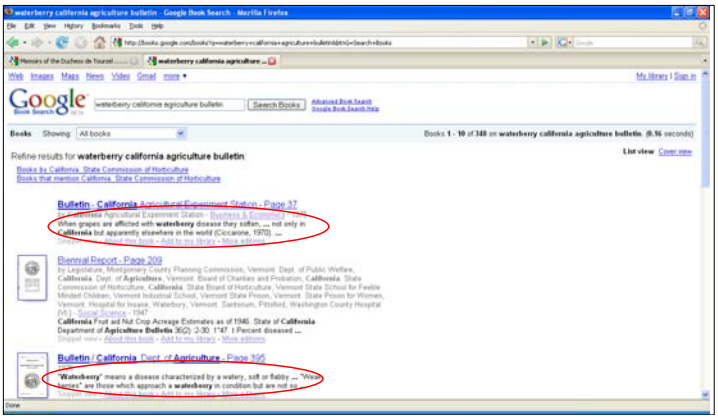
Why Engage in Mass Digitization?

- **Improve discovery and access**
  - indexing the full text of every book and making that full text available via Internet search engines makes our books easier to find by placing them where the users are.
  - Deliver books digitally wherever users are – laptops, mobile devices, ebook readers…
- **Preserve and protect our collections**
  - In earthquake and fire-prone California, digitizing books in our collections will protect the university from catastrophic loss should disaster someday strike our libraries
- **Enhance student and faculty research**
  - Scholars can trace the evolution of ideas and perform other sophisticated textual analysis, opening up new avenues of scholarship
- **Support collection management**
  - by making our collections more available digitally, we can explore more efficient and effective ways to manage our print collections
  - makes the 11 million volumes in our regional library facilities browsable
- **Fulfill our public service mission**
  - Many books of enduring general interest that are in the public domain – including classic works of literature but also more unique items such as early histories of the settlement of California and the West – can now be read by anyone, anywhere, anytime.

Use Cases



Discovery: A Berkeley environmental scientist found new historical material on waterberry using Google Book Search



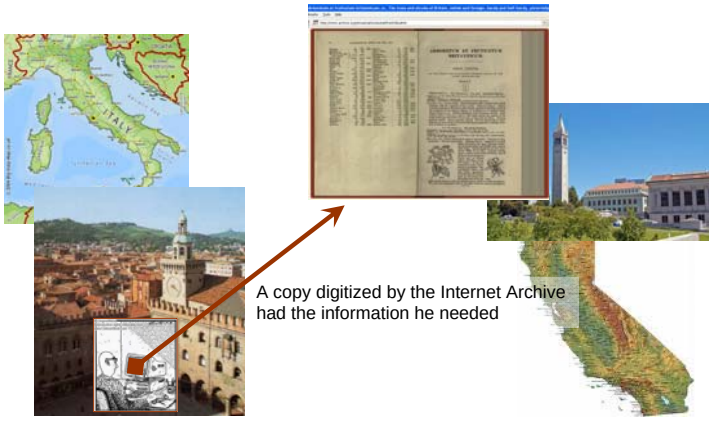
Discovery: Stories from the trenches

"I find it particularly helpful for newer material when I have a narrow topic and I'm not quite sure what type of book will cover the topic. Sometime people want information at the more general level of a book. Here I am primarily using info provided by the publisher. The scans of some of the journal literature have also been very useful."

"I remember a call from Canada from someone who had found an early description of a plant species in Google Books. Although the item was quite old, he could only see a snippet. I could see the entire page and read him the information he needed."

"There are times that I've searched for subjects for students papers and found scanned copies of Santa Cruz books that we have sitting in our stacks. Because of Santa Cruz's space problems, they have stored things that are used a fair bit at Berkeley. This has helped me help patrons identify useful books in the Bioscience Library stacks."

Access: An Italian scholar wanted to consult a Berkeley volume of Arboretum et Fruticetum Britannicum

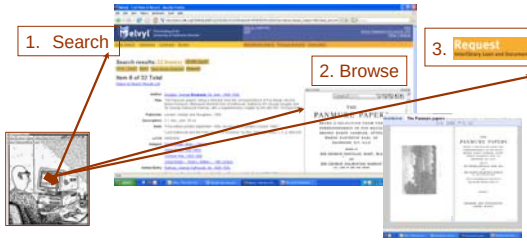


Access: Faculty and students can browse items at an RLF before requesting delivery

11 million volumes from UC library book collections are stored in our regional library facilities

"It is not an onerous process to look at something that has been moved offsite – you just fill out a form, and the requested item arrives in a day or two. This is an efficient system for many things, but not for browsing. Browsing does not work in two day intervals. It feels like playing chess by mail, a game that has never appealed to me."

Ammon Shea, *Is A Book In The Library Worth Two in the Offsite Storage Facility?*
http://blog.oup.com/2009/02/library_storage/

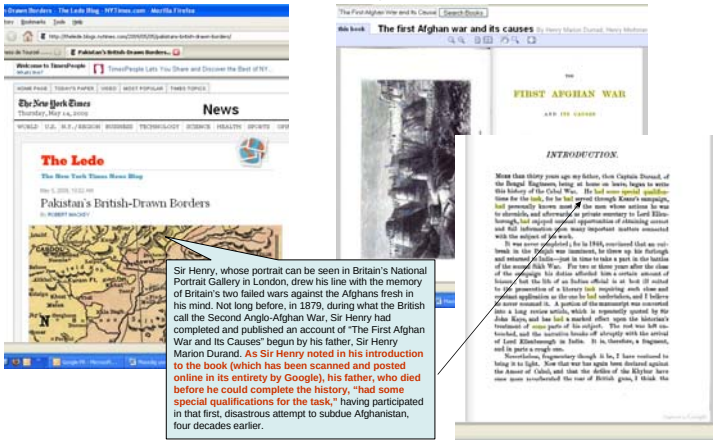


Southern Regional Library Facility, UCLA

Advanced research: Data mining and computational analysis

- Pending CARTA NSF grant proposal: "The Museum of Comparative Anthropogeny (MOCA): A Virtual Organization for Explaining the Origin of Humans".
  - "One of our tasks will be to mine digitized texts related to anthropogeny and glean themes, theme evolution, term usage & evolution, develop a basic thesaurus, etc. […] We would like to know which of these books is readily available in digital form …"

Public benefits: Mass digitized books are beginning to inform serious journalism



Sir Henry, whose portrait can be seen in Britain's National Portrait Gallery in London, drew his line with the memory of Britain's two failed wars against the Afghans fresh in his mind. Not long before, in 1879, during what the British call the Second Anglo-Afghan War, Sir Henry had completed and published an account of "The First Afghan War and Its Causes" begun by his father, Sir Henry Marion Durand. **As Sir Henry noted in his introduction to the book (which has been scanned and posted online in its entirety by Google), his father, who died before he could complete the history, "had some special qualifications for the task,"** having participated in that first, disastrous attempt to subdue Afghanistan, four decades earlier.

Benefits for Library Collection Managers

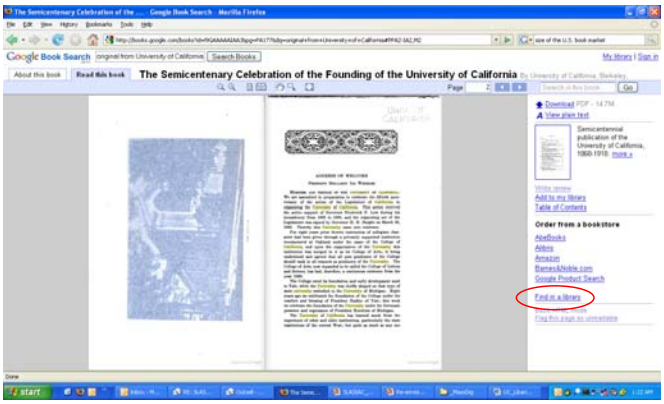
- Reclaim valuable library space
  - By reducing duplication across UC libraries By optimizing access to lesser-used volumes placed in more efficient remote storage
- Preserve and protect collections
  - From wear and tear
  - From incidental as well as catastrophic loss
- Reduce costs
  - Interlibrary loan, preservation
  - Less need to license e-versions of out-of-copyright material already in our collections
- Or, will mass digitization increase demand for our physical collections???
  - The physical book is still an incredibly efficient and satisfying technology, especially for long narrative text
  - Mass digitization does not replace the book as artifact

Where can you find UC books?

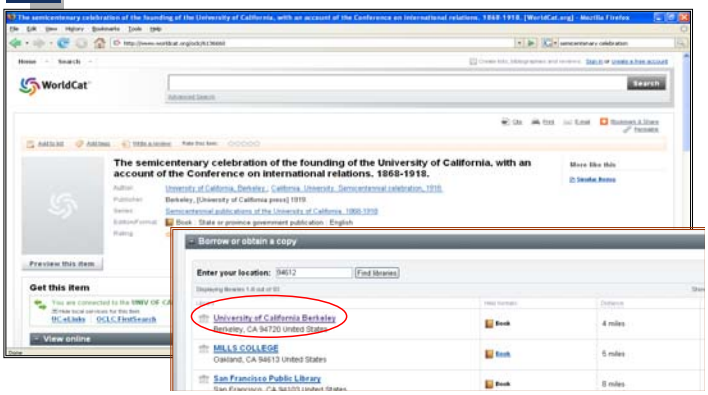
- **Google Book Search**: http://books.google.com/
- **Internet Archive**: http://www.archive.org/details/university_of_california_libraries
  - Full text searching across multiple texts is not currently supported by Internet Archive
- **Melvyl:** http://melvyl.cdlib.org/
  Next Generation Melvyl: http://melvyl.worldcat.org/
  - Currently only links to Google books are available
- ***Soon…*: HathiTrust** http://hathitrust.org

Google Book Search

Full text viewing in Google Book Search

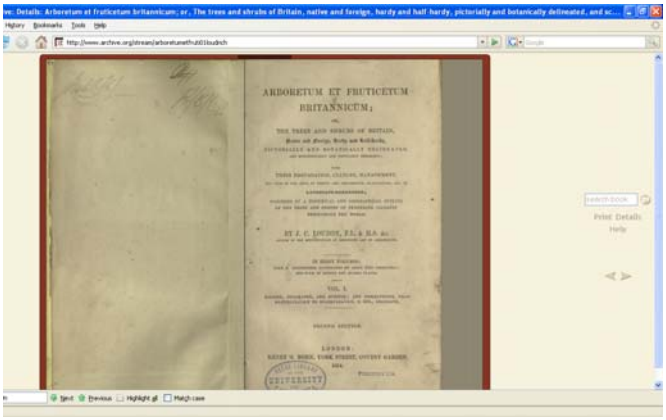


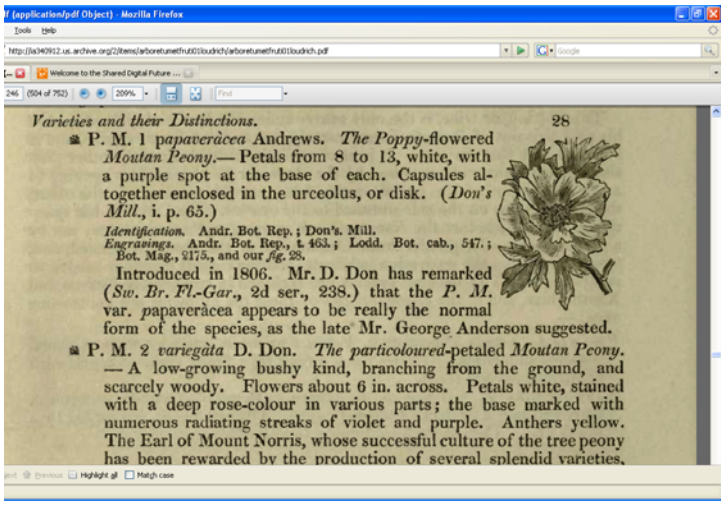
Google “Find in a Library”



Internet Archive Site

Full text at Internet Archive





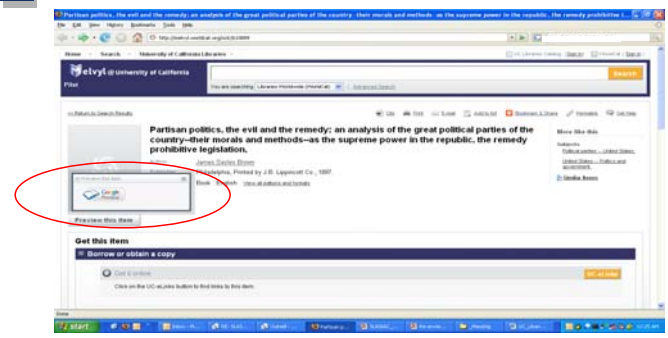
Google books in Melvyl

- Links are enabled via a Google API*
  - includes an embedded viewer application
- Limitations:
  - Currently links are only available for books with a standard identifier (ISBN, OCLC number, etc.)
    - Older works lacking identifiers are not yet accessible
  - Links to Internet Archive books are not available in UC catalogs at this time

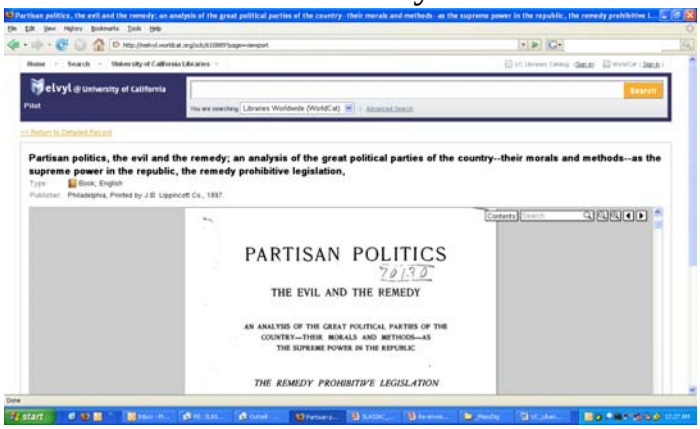
* API = Application Programming Interface

…embedded viewer in Melvyl

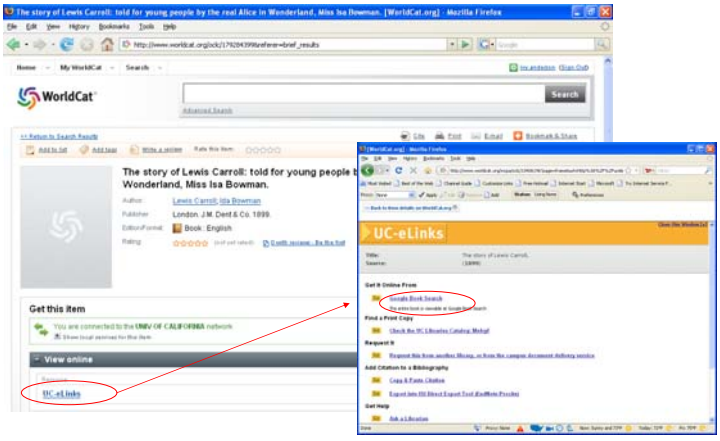
…in Next Generation Melvyl



…embedded viewer in NextGen Melvyl



… WorldCat access via UC-eLinks

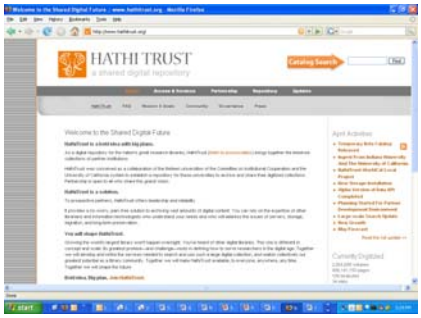


…and soon…

HATHI TRUST
a shared digital repository

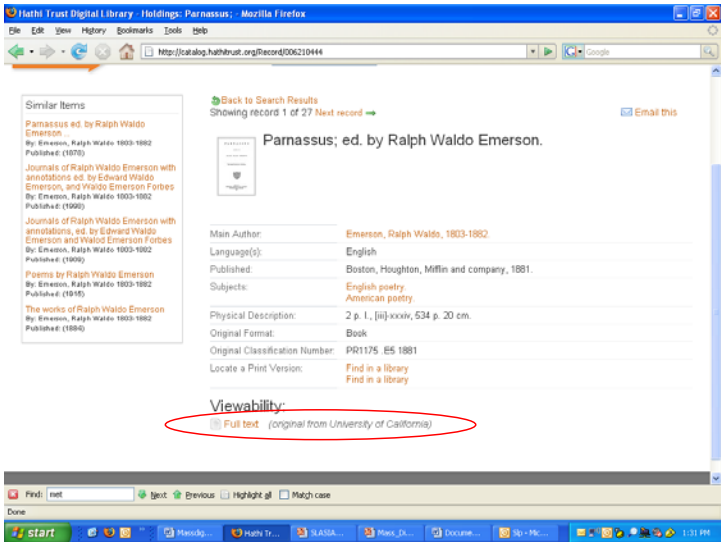
http://www.hathitrust.org

Currently Digitized
2,839,932 volumes
993,976,200 pages
106 terabytes
34 miles
2,307 tons
452,740 volumes (~16% of total)
in the public domain



What is the HathiTrust?

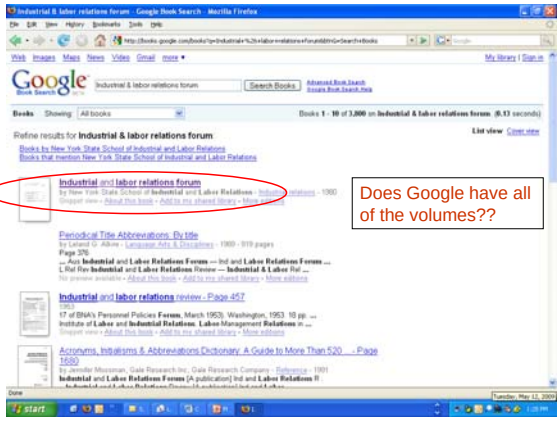
- A shared digital repository for mass digitized content, founded in October 2008
  - Operating at web scale
  - 2.8 million volumes now
  - Will be 5 million+ volumes with UC
- Members
  - CIC Libraries (Committee on Institutional Cooperation)
  - University of California
  - University of Virginia
  - More institutions may join in future
- Lead partners
  - University of Michigan
  - Indiana University
  - University of California
- Where are the digitized files stored?
  - Servers at the University of Michigan and Indiana University
  - Additional mirror sites may be developed in future

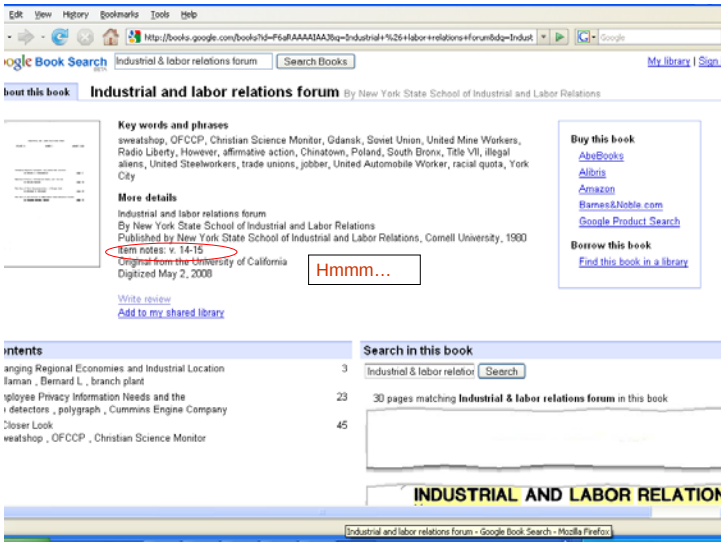


Why is UC participating in HathiTrust?

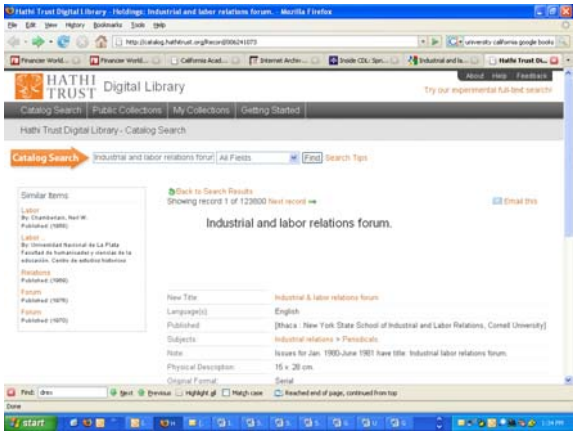
- Preservation and stewardship of UC resources
  - Brings our Google and Internet Archive books together in a common preservation repository under UC control
- Economy of scale
  - Storing mass digitized books is expensive – many terabytes of data
- Better access to our own books
  - Create robust links to full text in HathiTrust from Melvyl and WorldCat Local, including *all* viewable content from UC and other participating libraries
  - Build improved access interfaces via the HathiTrust API
- Aggregate multiple library collections for greater research impact
  - HathiTrust will support shared access and search mechanisms across all partner content to the extent possible
  - With UC, nearly 5 million books and counting
  - ¾ million books in the public domain
- Experiment with largescale search, text mining, and other specialized services developed with academic users in mind
  - Google and Internet Archive are building services for the general user
  - Research libraries will build services optimized for serious research

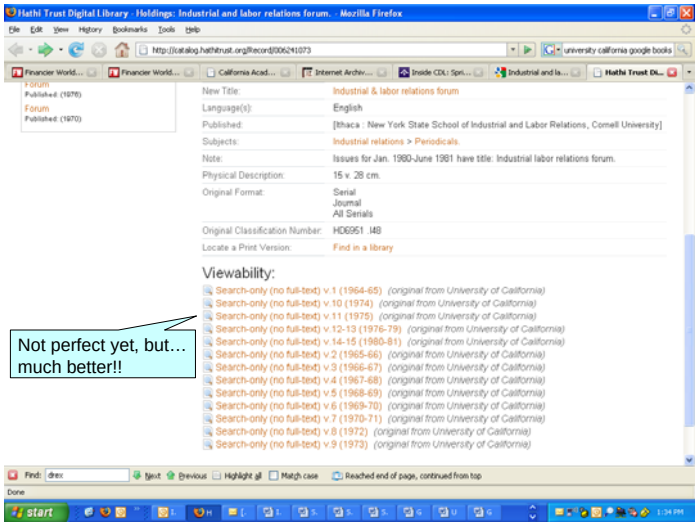
Improved services via HathiTrust: Multi-volume works





The same title in HathiTrust





When will HathiTrust be available?

- UC Google books are being ingested into HathiTrust over the next several months
- UC Internet Archive books will follow
- CDL is beginning to investigate access mechanisms in concert with Michigan and other HathiTrust partners
  - Planning discussions are underway with OCLC for a HathiTrust catalog based on WorldCat Local
  - UC will be able to add catalog links to all of its mass digitized books
  - "Collection builder" functionality will allow users to create and share specific curated collections
  - More advanced search and text mining to follow
- Building robust services will take time

How will all this be affected by the Google Settlement?

Google Settlement: Basic Facts

- In October 2008, Google settled a class action lawsuit brought by organizations representing authors and publishers, who claimed that Google's library scanning program violated their copyrights. Google has always claimed that this was fair use and legitimate under copyright law.
- The Settlement covers in-copyright books published prior to January 5, 2009. Public domain works, journals, and certain other categories of books are not covered.
- If approved, the Settlement will create a range of new services and business models that were not conceived at the time the library partner program was developed.
- The court cannot change the Settlement – it can only approve or disapprove it. At this time we do not know if the court will approve the Settlement. A court hearing is currently set for October 7th.
- *Impact on UC:* UC must amend its contract with Google to conform to the Settlement in order to retain copies of in-copyright books that Google has scanned.

Services Created by the Settlement

- **Public Access terminals** in public libraries across the country that will allow the general public to find and read books that are out of print or in the public domain
- **An Institutional Subscription** for access the full text of millions of out of print books digitized from libraries around the world
  - Books in the institutional subscription will have persistent links for use in electronic course reserves, course reading lists, etc.
- **A Limited Subscription** for free access to our own scanned books, should UC choose not to license the full Institutional Subscription
- **A Research Corpus** that will support advanced computational research on the full text of millions of books that Google has digitized
- **Services for visually-impaired users** to read and access all of the volumes Google has scanned
- **Consumer purchase models** allowing individual end-users to purchase online access to books.

Existing Google services will also remain available

- **Google Book Search** will continue to make the full text of all books searchable online
  - **"Find in a library"** pointers will lead users to the copies in libraries
  - More books in GBS will have 'preview' mode enabled, for better browsability
- **UC will receive copies of all of the books scanned from our collections**
  - Use of the digital files will depend on the copyright status of the book
  - At a minimum, we will be able to use the digital files to replace missing or deteriorated copies in our collection when needed
  - These copies will be stored in the HathiTrust shared repository
- **Books in the public domain can be used and downloaded freely** by scholars and the general public
  - Libraries can share their copies with other academic institutions for scholarship and research

What the Settlement won't allow

- There are a few things we won't be able to do with our own digital copies of the Google books
- *We cannot:*
  - allow full text viewing of in-copyright works
    - Text will be viewable through the limited or institutional subscription
  - Use in-copyright books for interlibrary loan or e-reserves
    - reserve links will be possible from the limited or institutional subscription
  - allow access via 3rd-party search engines and automated crawlers

Controversy over the Settlement

- The Google Settlement is not without controversy. Some are concerned that it will:
  - Give Google a monopoly over book digitization and suppress competition
  - Allow Google to charge high prices for subscriptions
  - Create an artificial market for orphan works, give Google a monopoly over them and prevent more open sharing of those works
    - *Orphan works = works still under copyright whose copyright owners cannot be identified or located*

Some responses to the Settlement controversy

- Giving Google a monopoly
  - The book market is large and diverse
    - Most book sales are for very current materials; out of print books are a minor factor
    - Many out of print works are in the public domain with no barriers to competition
      - Many libraries, including UC, will be offering print-on-demand services for their public domain books
    - Microsoft and others had the same opportunity, but withdrew
  - The Book Rights Registry and individual rights holders can strike deals with other providers, and are expected to do so
    - The Registry will have an incentive to work with other distribution channels to demonstrate its relevance to rights holders
    - Rights holders can also convey broader use rights than those defined in the Settlement

Some responses to the Settlement controversy

- Charging high prices for subscriptions
  - The Settlement's broad distribution requirement will work against this
  - Competition with other services is likely to keep prices reasonable
    - Free services: book search, find in a library, public access service, public domain works
    - Consumer purchase
  - Libraries are experienced in negotiating access to ebook content and assessing fair pricing
  - Pricing is subject to a formal challenge process including binding arbitration

Some responses to the Settlement controversy

- Locking up orphan works
  - As rights holders surface via the Settlement claims process, the true scope of orphan works will be better known
  - Once that happens, providing access to works that remain unclaimed will entail less risk than it does currently, making the legal protections offered to Google under the Settlement less of an advantage
  - Orphan works legislation will override the Settlement terms if enacted, making Google's privileged position short-lived

UC Libraries Assessment

- As one commentator has written:
  - "*The settlement is not what you would come up with if you began with a blank piece of paper and designed the optimal system for all the interested parties.*"
- Nonetheless, the Settlement will:
  - Make millions of books in research library collections more accessible to users and the general public than ever before – including more accessible than they are now via Google Book Search
  - Provide a significant corpus of material for advanced computational research
  - Allow individual rights holders to convey broader use rights if they wish
  - Allow UC to retain its copies of Google in-copyright scans for replacement purposes and for the creation of additional services
  - Create additional opportunities for management of print collections
  - Potentially, spur a more rational legislative solution for orphan works
- The Libraries are creating a web page and FAQ with more information about the Settlement and its implications for faculty

Finally…print is not going away

- *Books will always have artifactual value*
- *But…Mass Digitization will:*
  - Improve access to our collections
    - Internet-based discovery
    - Full-text searching and viewing
    - Data mining and computational research
    - Embedded links
  - Create collection management opportunities
    - Preserve and protect our collections
    - Reduce duplication among print collections
    - Enhance economical print storage with full-text browsing
    - Optimize use of valuable library space
  - Allow the UC Libraries to better understand our users' needs for print vs. digital collections
  - Help us shape the library of the 21st Century for the 21st Century user