

UC Libraries Mass Digitization Projects FAQ

Last Updated: 6/3/11

The Basics -

1. What is mass digitization?

The goal of mass digitization is to digitize the books in the world's libraries on a grand scale - ideally, every book ever printed. Millions of books from the UC Libraries will be scanned through our participation in mass digitization projects. To do this economically and with some speed, mass digitization is based on the efficient photographing of books, page-by-page, and subjecting those images to optical character recognition (OCR) software to produce searchable text. Human intervention is reduced to a minimum so the OCR output is generally used without undergoing additional revision. Also, only limited structural markup, such as page numbers, tables of contents, and indices, are included.

2. What are the University of California's mass digitization goals?

Mass digitization projects expand the UC Libraries ability to give faculty, students and the public access to information and support our exploration of new service models. These projects are designed to:

- Enhance student and faculty research. Mass digitization of these materials increases awareness of the rich materials in our collections and enhances access.
- Enable scholars to trace the evolution of ideas and perform other sophisticated textual analysis more easily by indexing the full text and making it searchable by computer, supporting scholarship in new ways.
- Fulfill its 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.
- Preserve and protect our collections - In earthquake and fire-prone California, digitizing the books in our collections may also help protect the university from catastrophic loss should disaster someday strike our libraries.

Mass digitization allows the UC Libraries to explore new questions and service models including but not limited to the following areas:

- Enhanced Discovery and Access: how will improved access to mass digitized materials in our print collections support the research, teaching, and private study needs of students, faculty, and other library users?
- Collection Management: can mass digitization help support our efforts to manage campus print collections and build more effective shared print collections?
- New Services to Users: what new service opportunities and/or research paradigms are enabled by massively digitizing our library collections?
- Curating through Collaboration: will participation in mass digitization projects help create access for our users to third-party materials not currently available through our own libraries?
- Funding Reallocation: to what extent can the digital reformatting of our own collections of public domain works obviate or lessen the need to allocate funds toward licensing online collections of these same materials from commercial providers?

3. What mass digitization projects are currently underway at UC?

The UC Libraries are currently participating in the Google Books project. This is a non-exclusive agreement and the UC Libraries may enter into other agreements with other digitization projects as they arise. Over the past few years, UC has also accomplished large-scale digitization in partnership with the Internet Archive.

Google Books: <http://books.google.com/>

In the Google Books project, books and serials (both in-copyright and in the public domain) in all languages are being scanned. The Google-University of California contract targets scanning 2.5 million volumes over a period of six years.

Internet Archive: <http://www.archive.org/details/americana> and <http://www.archive.org/details/texts>

Our work with the Internet Archive has been project driven, with funding provided by various organizations including the California Digital Library, Microsoft, Sloan Foundation, Kahle-Austin Foundation, Omidyar and Yahoo. UC Libraries' earliest forays into mass digitization were under the auspices of the Open Content Alliance (OCA), which represented the collaborative efforts of an international group of cultural, technology, governmental organizations, and nonprofit organizations including founding partner Internet Archive.

4. What books are being scanned?

Currently we are digitizing Books from the collections of UC Libraries, housed in the Northern Regional Library Facility (NRLF), and at UCLA, UC Santa Cruz, and UC San Diego.

The Scanning Process -

5. Who is doing the digitizing?

Google is scanning books and serials for the Google Books project. The Internet Archive has been the digitization agent for the Open Content Alliance, Microsoft-funded digitization, and other projects. Books are not destroyed during the digitization process.

6. How and where is the digitizing being done?

Books scanned through the Google Books project are being digitized offsite in a Google-managed facility. Books scanned by Internet Archive were digitized at Internet Archive facilities including the SRLF facility within the UC Libraries system.

7. What will happen to the books after digitization?

All books are returned to their home locations after digitization. Books are generally returned to the shelf within two to three weeks.

8. Are there standards regarding the quality of the scans?

CDL has been engaged with technical staff at Google and the Internet Archive regarding quality standards for the two scanning projects. Partner Libraries for the mass digitization projects have developed technical specifications for image compression to ensure efficient and high-quality long-term storage of the derived page images.

9. What rights to the digitized content does UC have in the projects; is access limited in any way?

All contracts specify that UC digital images will be available to the UC Libraries to download and manage. The UC Libraries' digital copy is subject to certain rights and restrictions regarding use and distribution. The University of California's use or ability to display the downloaded copies of the full text of all books is subject to the restrictions of copyright law. Full-text searching will be possible for all of the digitized books, but some scanned books will not be completely viewable due to copyright restrictions. Specifics include:

Google

- UC Libraries have the right to use the UC Libraries digital copy at the University's sole discretion, subject to copyright law, as part of the services offered to University Library patrons (including all individuals and organizations served from the UC Libraries websites).
- UC Libraries must implement technological measures to restrict automated access by crawlers, robots, spiders etc. to the UC Libraries digital copy.
- UC Libraries may not permit downloading for commercial purposes.
- UC Libraries may not knowingly permit the automated downloading and redistribution of the UC Library digital copy by third parties. UC Libraries must develop methods for ensuring that substantial portions of the UC Libraries digital copies are not downloaded from the UC Libraries website or otherwise disseminated in bulk.
- UC Libraries are permitted to distribute no more than 10% of the UC Libraries digital copy to other libraries and educational institutions for non-commercial, research, scholarly, or academic purposes (but not any portion of image coordinates).
- UC Libraries are permitted to distribute all or any portion of public domain works contained in the UC Libraries digital copy (but not any portion of image coordinates) to other research libraries for use by those libraries' authorized students, faculty, and staff for research, scholarly, or academic purposes.
- Image coordinates, which link words in the OCR'd full text to specific locations on the viewable page, may not be shared with any entity.

Internet Archive

- There are no restrictions on access or redistribution placed on the UC Libraries' digital copy.

10. How can our patrons access these texts, i.e. through Melvyl®, or local catalogs, or a webpage, any search engine, or....?

UC Libraries patrons can currently access UC Libraries scanned books in a number of different locations:

Melvyl: <http://melvyl.worldcat.org/>

(Links to HathiTrust and Google books in UC and all partner collections)

HathiTrust: <http://catalog.hathitrust.org>

Google Book Search: <http://books.google.com>

(Currently there is no ability to browse or search the UC Libraries subset)

Internet Archive: http://www.archive.org/details/university_of_california_libraries

For more information, please see: [Where to Find Our Books](#)

11. Will access be different for the general public than it is for our faculty, staff and students?

The general public and UC faculty, staff and students currently have the same access to digitized UC Libraries books when searching and browsing Melvyl and the websites listed below:

[HathiTrust](#)

[Google Book Search](#)

[Internet Archive](#) and [Open Library](#)

and [Melvyl](#)

12. How are patrons be impacted while books are being digitized?

Patrons are be impacted as minimally as possible during the mass digitization projects. Books are usually returned to the shelves within a few weeks or less.