Digitization of Library Collections

Joy Paulson
Digital Collections Librarian
Mann Library, Cornell University
Digitization Initiatives at Mann Library

• Cornell University Library began to digitize collections in 1991
• Mann Library began digitizing collections in 1995
  – Focus on core historic materials
  – Microsoft Book Project
  – Google Book Project
CHLA (Core Historic Literature of Agriculture)
HEARTH (Home Economics Literature: Research, Tradition, and History)
Microsoft Book Project

- Cornell University Library partnered with Microsoft to digitize materials from its collections
  - Single volume monographs
  - English language
  - Public domain titles only
  - Mann Library contributed 12,000 volumes
  - Available on the Internet Archive (www.archive.org)
Google Book Project

• Cornell University Library has been a Google Library partner since 2008.
  – Minimum of 500,000 volumes from all subject areas in Cornell’s collections
  – Monographs and Serials
  – All languages
  – All date ranges

• Images sent to HathiTrust (http://www.hathitrust.org)
Google Book Project

- Mann Library participation
  - 2008-2010
  - Agriculture, life sciences, human ecology, ornithology, entomology, experiment station materials
  - Over 250,000 volumes
  - All titles searchable in Google Books and public domain materials are fully viewable
  - Mann’s images are in HathiTrust
Selection for Digitization

- What is digitization?
- Why digitize?
- What should be digitized?
- Copyright considerations
What is digitization?

Digital images are electronic snapshots taken of a scene or scanned from documents, such as photographs, manuscripts, printed texts, and artwork.
Digitization converts materials from formats that can be read by people (analog) to a format that can be read only by machines (digital). Flatbed scanners, digital cameras, overhead digital camera workstations, planetary cameras, and a number of other devices can be used to digitize library.
Why Digitize?

- To create greater access to library materials
  - Collections are accessible 24/7
  - Full-text searching
  - Ability to present information in new and important ways
  - Gives access to library materials to those who have lacked access in the past
- Improve preservation
Scan from a glass plate negative of a pioneer sod house in Nebraska
Preservation

Digitization can also help preserve precious materials. Making high-quality digital images available electronically can reduce wear and tear on fragile items if the use of the original paper materials is reduced.
Selection for Digitization

• Many approaches to selection for digitization
  – Specific collections
  – Core literature approach
  – Mass digitization
What should be digitized?

• Rare and special collections materials
  – Manuscripts
  – Rare books
  – Photographs
  – Maps
  – Archival material

• General collections materials
  – Monographs
  – Serials
• Specific types of material to be digitized
  – Local, regional, and national agriculture journals
  – Extension materials
  – Historical materials relating to agriculture
  – Unique materials
  – Government publications
  – Census material related to agriculture and rural populations
Copyright Considerations

Copyright is a complex issue that strongly impacts the selection of materials for digitization. Many libraries and museums have chosen to avoid the complexities of copyright law by digitizing materials that have passed into the public domain and are no longer covered by copyright restrictions.
• But limiting digital collections to material not under copyright protection means many valuable materials are not available online.

• Mann Library, especially when digitizing serials, has worked with copyright holders to gain permission to put historical materials still under copyright protection online.
### Rural Sociology

The image shows a list of issues from the *Historical Literature of Agriculture* (HLA) database, specifically the Rural Sociology section. The issues listed are:

- 1932 (v 1 n 1 - v 1 n 4)
- 1934 (v 2 n 1 - v 2 n 4)
- 1936 (v 3 n 1 - v 3 n 4)
- 1938 (v 4 n 1 - v 4 n 4)
- 1941 (v 5 n 1 - v 5 n 4)
- 1944 (v 6 n 1 - v 6 n 4)
- 1947 (v 7 n 1 - v 7 n 4)
- 1949 (v 8 n 1 - v 8 n 4)
- 1951 (v 9 n 1 - v 9 n 4)
- 1952 (v 10 n 1 - v 10 n 4)
- 1953 (v 11 n 1 - v 11 n 4)
- 1955 (v 12 n 1 - v 12 n 4)
- 1957 (v 13 n 1 - v 13 n 4)
- 1959 (v 14 n 1 - v 14 n 4)
- 1961 (v 15 n 1 - v 15 n 4)
- 1963 (v 16 n 1 - v 16 n 4)

Title from journal information screen (viewed May 2004)

Digitization funded by Cornell University Class of 1956. Title selected from the series Literature of the agricultural sciences for the Core historical literature of agriculture.
Digitization Standards

- In-House versus Outsourced Digitization
- Formats
- Resolution
- Bitonal/Grayscale/Color
- Metadata
- OCR
In-House versus Outsourced Digitization

- Deciding whether or not to digitize in-house or hire an outside company to do the work is a major decision, and largely depends on the materials to be digitized.

- Outside vendors are generally very efficient at digitizing materials of similar format and size. However, it may be necessary to familiarize them with library standards and how to handle fragile material.
• Creating the capacity to scan rare and extremely fragile material in-house may be necessary, but involves an investment in scanning equipment, IT hardware, and staff expertise.
Formats

• Master Images
  – Important to use standardized formats
  – TIFFs
  – JPEG2000

• Display Images
  – JPEGs
  – PDFs
  – Other formats--GIFs
Resolution (master images)

- Black and white/Bitonal
  - 600 dpi/ppi
  - 1 bit
- Gray-scale
  - 300-400 dpi/ppi
  - 8 bit
- Color
  - 300-400 dpi/ppi
  - 24 bit
Metadata

• Metadata is usual defined as “data about data”

• Types
  – Bibliographic
  – Technical
  – Structural
  – Rights management

• Metadata schema—METS, Dublin Core
Structural Metadata

• Monographs
  – Title page
  – Table of contents
  – Tables of illustrations, maps, etc.
  – Indexes
  – Illustrations
• Serials
  – Title page
  – Tables of content, illustrations, maps
  – Indexes
  – Illustrations

• Article level metadata
  – Author of article
  – Title of article
  – Page range
Optical Character Recognition (OCR)

- For printed collections OCR of the text allows full-text searching
- Quality of OCR is related to the complexity of the page layout--pages with multiple columns and multiple fonts make it more difficult to obtain high quality OCR
- Quality of the print in the original also effects OCR quality
• Quality of the dictionary for the language being OCR’d also affects quality
• Hand-written archival materials and manuscripts cannot be OCR’d at this time.
Online Access Considerations

- Platforms
- Browsers
- Expertise to manage and maintain collections
Platforms

- Greenstone
- Content dm
- Fedora
- DLXS
- Many others
Browsers

• Websites need to work on a variety of browsers
  – Internet Explorer
  – Mozilla FireFox
  – Google Chrome
Expertise to Manage and Maintain Online Collections

• IT staff
  – Website design
  – Bring collections online (images, metadata, OCR)
  – Maintain servers
  – Maintain long-term archival repositories
Resources

• Moving Theory Into Practice: Digital Imaging Tutorial, Department of Preservation, Cornell University Library: http://www.library.cornell.edu/preservation/tutorial/index.html

• Anne Kenney and Oya Reiger, Moving Theory Into Practice: Digital Imaging for Libraries and Archives, Research Libraries Group (RLG), 2000
• A Framework for Building Good Digital Collections, Institute of Museum and Library Services

1. Assuming an ideal world where you faced no resource constraints, identify print resources you would like to see digitized.

2. What internal and external constraints would make it difficult to reach this goal?

3. What resources would be necessary for your institution to participate in a digitization project to digitize material for higher education for agriculture?

4. List several ways your institution could contribute to a national effort to digitize agricultural material.