Old Wine in New Wineskins: Sustaining Access to and Preserving Legacy Digital Collections

Joy Paulson
Cornell University

Core Historical Literature of Agriculture

- Materials scanned as part of R&D projects in the early 1990s.
- Department of Education Title II C grant in mid-90s.
- NEH grant in mid-1990s.
- 6 serial titles digitized with funds from a private gift in 2000-1

CHLA History

- The collection was available on a number of different platforms in the mid- to late 90s.
- 1999: the collection was not available online.
- 1999 – 2001: Project Prism
- 2000-2001: 816 monographic volumes were made available, but the rest of the collection need a substantial amount of work.
- 2004: Reworked the project

The Good News

- All files were 600 dpi bitonal TIFFS
- Only a small number of files were corrupt

The Bad News

- Gathering the files together
- Unresolved copyright issues
- Metadata
- OCR

Why would we choose to do this?

- This collection was considered extremely important.
- When available online the materials received a high level of use.
- We wanted to continue to add titles to the collection.
- Increased ability to preserve collection

Gathering the Files

- CHLA materials currently online.
- Materials that had been made available as part of other collections.
- R&D materials
- Serials
- A few surprises

Copyright: the never ending problem

Bibliographic Metadata

- Updated the cataloging to current standards.
- Discovered that for a few titles the electronic version had never been cataloged.
- Created links from the catalog directly to the title rather than just to the collection.
- Transferred the bibliographic metadata to the project programmer.

Structural Metadata and OCR

- Restructured all titles for the collection
- Evaluated and compared the OCR created using a couple of different OCR engines.
- OCR'd any files that hadn't been OCR'd.
- Ensured that the images were all good.

IT Process

- Platform: DLXS
- Gathering the files
- Pulling everything together
- Surprising problems
  - Sideways images
  - Images that couldn't be converted on the fly
  - Problems viewing PDF images

Costs

- Copyright: 125 hours
- IT: 400 hours
- Structural Metadata: 700 hours
  - 300 support staff hours
  - 400 student assistant hours
- Bibliographic metadata: 450 hours
  - 325 support staff hours
  - 125 librarian hours

- 125 project management hours
- TOTAL: 1,800 hours

CHLA Today (http://chla.library.cornell.edu)

- 1,834 monographic volumes
- 6 serial titles in 288 volumes
- 840,344 pages
- Small number of titles ready to be added to the collection.
- Small number of titles with unresolved copyright issues.

Future Plans

- Will add another 100,000 newly scanned images to the collection in the next 12 months.
- Add color images
- Will continue to slowly add to the collection with internal funds.
- Seeking grant funds to complete the project.

Core Literature Group

- Joy Paulson, Preservation Librarian
- France Webb, Programmer/Analyst
- Nathan Rupp, Metadata Librarian
- Keith Jenkins, Metadata Librarian
- Roswitha Clark, Technical Services Assistant
- Brian Lowe, Technical Services Assistant
- Johanna Hartnagel, Preservation Assistant
- Barbara DiSalvo, Special Projects Librarian