Creating Digital Text Collections

Using CONTENTdm
Introduction

- Types
- Special Considerations
- Issues & Examples
- CONTENTdm features
Types of Digital Texts

- **Text**
  - Any source with written pages

- **Digital text possibilities**
  - Image only
  - Image and hidden text
  - Image and visible text
  - Encoded text
Special Considerations: Selection

- Uniqueness
- Utility
- Avoid duplication
- Homogeneity
Special Considerations: Delivery

- Target audience
  - General users
  - Scholars

- What do they need from original?
  - Structure
  - Information
  - Reproduction
  - Facsimile
  - Searchability
Image

● Is page reproduction necessary?
  – Image only
  – Text only (images only for OCR)

● Scanning Options
  – Black & white (line drawings or text only)
  – Grayscale (text with b&w images)
  – Color (text with color images; scholarly substitute)
Text

- Transcription
  - Outsource
  - Student workers

- Optical Character Recognition (OCR)
  - Scan page of text
  - Use software to read and convert text in image
  - Output text file
  - Visible (corrected) or hidden (dirty)

- PDF
Examples

- Image only file (without digital text)
- Corrected plain text file (without images)
- Corrected text used for HTML
- Corrected text with XML & page images
- Uncorrected hidden text with page images
- Color scholarly substitute
- PDF
CONTENTdm Text Options

- Compound object feature
  - Document
    - Sequential pages
  - Monograph
    - Documents with structure (chapters or sections)
- Images can be TIFFs, JPEGS
- Text files created separately or using OCR extension
- PDF options
Steps for Creating Files

- LOUIS creates collection
- Create database or spreadsheet
  - Track original names and object file names, and structure
  - Follow rules for tab-delimited text files
- Scan pages and save as archival TIFFs in one directory
- OCR on images and save text files with same names (.txt)
Establishing Directory

- Tab-delimited text file with compound object metadata, and optional structural and page metadata (same name as directory)

- “Scans” sub-directory for images

- “Transcripts” sub-directory for text files (same names as page scans)
Importing into CONTENTdm

- Use correct wizard
  - Compound Object for individual files or test settings
  - Multiple Compound Object wizard for a whole series or more complex uploads
- Structural and page metadata in tab-delimited text file
- Save settings
- Preview
CONTENTdm PDF Options

- Single file, regardless of number of pages
- Can be part of a compound object
- Text extracted when approved and added to collection
  - Empty full-text search field required
  - First 128,000 characters
- Can be generated with OCR extension (not currently available in LOUIS)
  - Can generate print-copy of entire PDF
Pilot Project

- Acts of the Louisiana Legislature (training server)
  - Images with text transcriptions
  - PDF (individual pages)
  - PDF (pages grouped by section)
Conclusions

- Academic scholarship thrives on access to texts
  - Primary resources
  - Scholarly literature
- Many options available for creating text collections
- LOUISiana Digital Library needs to expand into this realm
Bibliography