Digital Collections, Repositories, and Archives

Greg Zick, Vice President
Digital Collection Services

ARL Research Library Leadership Fellows

November 1st, 2007
Future Direction
- New OCLC Strategy

Enterprise product architecture

WorldCat Grid

Local

Global

Group
Presentation Outline

- Digital Collection Creation
- Institutional Repositories
- Digital Repositories
- OCLC Future Directions
One way to digitize all the world’s collections

- List all the institutions that have collections?
  - Institutional Registry
- List all the collections at each institution
  - Archive Grid
- Digitize the items in the collections
  - Digital Registry
- Create the Digital Collection
  - CONTENTdm
- Archive original TIFFS
  - Digital Archive
- Add to Global Digital Repository
  - Harvest metadata into WorldCat
Digital Collection Creation

- CONTENTdm

© 1997-2005 DiMeMa, Inc.
Assess users’ needs and collections’ conditions

Convert your materials to digital collections and create metadata

Organize and store your digital collections

Expose your digital collections while providing Web access
What is CONTENTdm?

- A complete software solution developed by DiMeMa
- Store, manage and access your digital collections
- For organizations of all types and sizes from academic libraries, historical societies, public libraries and other collaborators
- Showcase a wide range of media types from photos and documents to audio and video files
Building partnerships and CONTENTdm digital collections

- Multi-regional or statewide consortia:
  AK, CA, FL, IA, IL, IN, NE, MN, MO, NY, OH, PA, WI
- Mountain West Digital Library – 20 members
- Western Waters Digital Library - 30 members
- Columbia River Basin Ethnic History Archive
- Digital Library of Appalachia - 12 members
- Historically Black Colleges and Universities Library Alliance
- Upper Mississippi Valley Digital Image Archive
- REALIA Project – Assoc. Colleges of the Midwest, Assoc. Colleges of the South and the Great Lakes College Assoc.
- Consortium of Liberal Arts Colleges (CLAC) – 62 members

These and many more...
Images: Photographs, posters, postcards
Colton’s Map of California, Nevada, Utah, Colorado, Arizona, New Mexico (1864)
OCLC Online Computer Library Center

Zoom and Pan
Dec. 13, 1902

Dear Sir:

Replying to your letter of the 11th inst. beg to say that I have no manuscript referred to, and that the preparation of a manuscript of the plots of abduction and assassination of Pres. Lincoln by John Wilkes Booth and others, and the escape of Booth, in a complete and authenticated way will require the expenditure of time, labor and money, and before beginning the undertaking should require that we

Transcribed text becomes searchable!
 CONTENTdm’s Integrated OCR Capability
Bureau of Economic Analysis

OCLC Preservation Service Center completed scanning and provided digital images to BEA.

Use of CONTENTdm to search and access.
HERRINGTON

KIM QUEY TELLS ON HERRINGTON
Confesses to Dr. Gardner That She Was Coached to Commit Perjury.

Sensation in the Case of the Mayfield Chinese Woman Who Was Delivered Into the Hands of the Highbinders.

ATTOREY J. N. E. WILSON, who appeared yesterday in United States Commissioner Haascock's court representing Bert Herrington in the case of the United States vs. Kim Quey, was buried clean out of the case by his own conduct. His client turned the tables on the gang of highbinders that were fighting for the possession of her
tself as a chattel, and the ring was filled with consternation. If her story is true one lawyer may be disbarred.

Kim Quey was arrested at Mayfield by Chinese Inspector Gardner for being illegally in the country, he having discovered that she was one of the show women who had come here to take part in the exodus.

According to her statement they said: "You must testify in order to clear us, for if you don't our case wouldn't be worth anything. We would be hanged." You must say that you were born in San Francisco, that you were not well and died in the Chinese Mission Home; this

FRANCISCO CALL, TUESDAY, APRIL 29, 1904.

TEN LESSONS IN FRENCH CONVERSATION.

ENGLISH-CHINESE.

COLUMBIA.

PUBLISHED EVERY SATURDAY.

ENCINO LETTERS.

COLUMBIA PRESS.

![ship](image_url)
Audio/video

Search Results: Your search retrieved 3 matching item(s).

Click on the thumbnail or title to view the item and corresponding description.

1. Elof Norman Interview
2. George Wardell Interview
3. Jack Leonard Miller Interview

Title: Jack Leonard Miller Interview
Location: United States--Washington (State)
Date: 
Photographer: 
Subject: 
Notes: Jack Leonard Miller (1899 - 1966) was one of the 74 Wobblies arrested following the Everett Massacre and charged with killing deputy Jefferson Beard. An articulate spokesman for the I. W. W., Jack was interviewed many times for film, video, radio and written histories. Hear an edited segment from an interview done by writer Jim Casey, at that time a columnist for the Everett Herald.

Original Format: Cassette Tape Interview
Repository: Everett Public Library
Order Number: 
Rights & Permissions: Copyright Everett Public Library 2002

Add to My Favorites  View My Favorites
<table>
<thead>
<tr>
<th>Image</th>
<th>Title:</th>
<th>Subject:</th>
<th>Description:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;Home Sweet Home&quot;, Peru, Indiana</td>
<td>Elephants; Camels; Horses; Baby animals; Circus animals; Drugstores; Banks; Streets</td>
<td>A line of circus animals parade down a street in Peru, Indiana. Adult elephants lead the way followed by a baby elephant and a group of camels. There is also a horse off to the side of the road. The animals pass various businesses including Agar Drug Company and the First National Bank. Several people have stopped on the sidewalk to watch the animals walk by.</td>
</tr>
<tr>
<td>2</td>
<td>&quot;The Squad 1907&quot; Indiana University, Bloomington, Indiana</td>
<td>Football; Sports; Sport clothes</td>
<td>Indiana University's 1907 football squad is shown. Most are wearing knee-length pants and pads.</td>
</tr>
<tr>
<td>3</td>
<td>&quot;Vandalia' Grounds, North Manchester, Indiana</td>
<td>Railroad locomotives; Railroad crossings; Railroad tracks; Automobiles</td>
<td>View of a railroad crossing in North Manchester, Indiana. A locomotive is shown approaching an intersection. Two lines of tracks cross the street. A car waits along the side of the road for the train to pass. Buildings are visible in the background.</td>
</tr>
<tr>
<td>4</td>
<td>#1 Anderson, Indiana</td>
<td>Bridges; Floods; Roads; Carriages &amp; coaches</td>
<td>Shows people standing by a bridge that crosses high water in Anderson, Indiana. The water appears to be nearing the level of the bridge, and is high up on trees that rise above the bridge. A horse and buggy are standing in the foreground, the road appears to be brick.</td>
</tr>
</tbody>
</table>
Discovery: WorldCat harvesting & Open WorldCat
One system, many solutions

CONTENTdm provides comprehensive solutions for your historical digital archives

Photographs, diaries, slides, newspapers, books, letters, audio and video oral histories, maps – all your digital assets!

http://www.contentdm.com/customers/
CONTENTdm 4.3

- Connexion digital import
  - New option for OCLC catalogers
  - Add digital resources to CONTENTdm via Connexion cataloging process
- Benefit
  - More options for building digital collections
  - Catalog using most convenient process for the organization
CONTENTdm 4.3

Connexion digital import

- Add items to CONTENTdm via the Connexion Client
- Digital collection growth built into cataloging workflow
- WorldCat MARC record crosswalked to Qualified Dublin Core and added to CONTENTdm
  - OCLC number stored in CONTENTdm – a global persistent identifier
- Digital items accessible by FirstSearch, WorldCat.org and WorldCat Local
- Requires OCLC Cataloging subscription, CONTENTdm license and CONTENTdm Hosting Services
CONTENTdm 4.3

Connexion digital import

- Metadata choices for cataloging
  - Connexion client (MARC)
  - CONTENTdm (DC, QDC, VRA)
    - Acquisition Station
    - Web-based Add option

- Serials support
  - Use “Attach Digital Object” in Connexion client for each issue in a serial item
  - 856 link will automatically retrieve a search results page with links to each issue stored in CONTENTdm
CONTENTdm 4.3

Connexion digital import

- In Connexion Client:
  - Attach Digital Content to existing record
  - Select CONTENTdm collection
  - Select file(s) from local computer/network
  - Replace command

- System processes metadata and file for import into CONTENTdm
  - Digital item sent to CONTENTdm collection
  - MARC metadata mapped to Qualified Dublin Core
  - Compound object creation, JPEG2000 conversion, and OCR or PDF processing, if applicable
  - Thumbnails generated
  - Link added to 856 field in WorldCat record
CONTENTdm 4.3

- Improved PDF handling
  - Convert multiple-page PDF files to CONTENTdm compound objects
    - Subset print options
  - Search term highlighting within PDF files
  - Creation of thumbnail images from PDF files
  - Improved text extraction from PDF files

- Benefits
  - More efficient processing of PDF documents
  - Single database of **ALL** digital resources
  - Better end user experience
CONTENTdm 4.3

- PDF files can be imported using standard options
  - Single or batch import via Acquisition Station
  - Web-based Add option
  - Connexion digital import

- Thumbnail images are automatically generated from the PDF when the item is added to the collection

- Text is extracted from the PDF and inserted into the full text search field when the item is added to a collection
  - Collection must have full text search field and that field must be empty when PDF is added to the collection
CONTENTdm 4.3

Compound object conversion

- When compound object conversion is enabled, CONTENTdm:
  - Creates a compound object based on the page order of the PDF.
  - Generates a page-level metadata record for each page.
  - Extracts text from the PDF, converts it to UTF-8, and inserts it into the full text field of the associated page level record.
  - Generates thumbnail images of each page of the PDF. The thumbnail image of the first page will also be used for the compound object.
  - Retains the original PDF file for export and printing.
  - Displays the PDF compound object in a compound object viewer with each page of the PDF accessible from the left navigation menu.
  - Highlights search terms in the PDF.
  - Provides an option to select a subset of the PDF to print or save.
Compound object conversion

approve items

Controlled vocabulary
All terms in pending items conform to the defined controlled vocabulary.

Review 4 pending items
date: 2000-01-01

<table>
<thead>
<tr>
<th>Item</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Custom Queries and Results</td>
<td>edit</td>
</tr>
<tr>
<td>Managing User Rights</td>
<td>edit</td>
</tr>
<tr>
<td>Full Text Searching</td>
<td>edit</td>
</tr>
<tr>
<td>Editing the Web Template Files</td>
<td>edit</td>
</tr>
</tbody>
</table>

delete all | approve all
Compound object conversion
PDF Enhancements

Printing and downloading

- Complete print version
  - Original PDF file retained for printing and saving
- Subset of print version
  - Select a subset of pages from the PDF to view, save, or print
  - Select all pages with search hits or pick individual pages or page ranges
  - Do not have to wait for large download if only need a few pages
  - Also available for non-PDF compound objects when they have been processed using the OCR Extension
Creating Custom Queries and Results

This guide details the use of Custom Queries and Results (CQR) in CONTENTdm 4. In order to use customized queries, you must have a working knowledge of HTML, and the ability to edit and create Web pages on your CONTENTdm server. If you do not have the ability to create Web pages on your server, please see your CONTENTdm system administrator to gain access.

Learn About

Section I: What is Custom Queries and Results?

Section II: Starting Custom Queries and Results

Section III: Using Custom Queries and Results

Custom Queries and Results or CQR enables the creation of custom Web interfaces and search options for CONTENTdm collections. Use CQR to specify how your collection should be displayed and generate HTML code that can be pasted into the Web pages used to display your collection.

CQR is not an HTML editor. It is a tool that uses your specifications to generate HTML segments that define queries of collections. Copy and paste the HTML into your Web pages using your preferred HTML editor. Using CQR, complex queries are simple to formulate and add to Web pages that interface with CONTENTdm. A working knowledge of HTML and building Web pages is required to interpret the query code into your collection pages.

Using Custom Queries and Results you can define five types of queries to CONTENTdm collections:

- Single Hyperlink: Create a single hyperlink to specify a single query. Because CONTENTdm is Web-based, all queries to collections have an associated URL that can be linked from your Web pages.
CONTENTdm 4.3

Compound object conversion

- Reduce the size of file that is downloaded for viewing
  - An entire PDF may be several MB but individual pages are much smaller
  - View a page within large PDF without downloading the full document
  - Increase speed of access to view
- Provide full text indexing by page not document
  - No secondary search required to find specific content in PDF
- Print only the information you need
- Better end-user experience!
CONTENTdm 4.3

Compound object conversion

- Quick and efficient for collection builders!
- PDF pages of compound object do not count against total number of items on the server!
- Ideal for born digital documents!
  - Theses, dissertations, government documents, e-publications, and more...
Digital Repository - future directions

- Automatic WorldCat registration and links
  - Register and replicate
  - Web-based user controlled mapping DC, MARC, VRA
  - Auto add WorldCat record #
- Seamless/transparent WorldCat ingest
- Update of all digital items with unique WorldCat reference number
- Integration with all features and functions of WorldCat.org
- Local branding and customization
OCLC Enterprise Strategy
“Discovery to Delivery” Collection Curation

- Consumer Platform
  - WorldCat.org “Indiana Cotton Mills”
- Management Systems
  - WorldCat Local “Hegg Alaska Photographs”, “meed”
- Content Platform
  - Digital Repository
  - Digital Archive
- Network Services
  - Use statistics
  - Collection Analysis
Digital Collection Services

**Digital Archive – re-engineered**

- Development of storage infrastructure in Dublin
- Redefinition and packaging of Digital Archive for market
  - long term archiving of digital masters
  - opportunity to reprocess as software improves
    - example OCR
- **Significantly reduce cost of service**
- **Two modes**
  - With repository
  - Mirroring
Digital Archive workflow for other systems
Type A service: batch mirror

From external content management system

Verify

Check-in

- Vol→ID→Checkin fixity, virus, format verification (JHOVE), etc.
- Verify manifest & files
- OK to Move ARCHIVE_ID

Copy to staging

Ingest

Volume Staging

Archival Storage

manifest

Metadata

Master Files

OK to Move

Copy to staging

OK to Move
Digital Repositories
cooperative advantage

- Broaden use of the repository
  - Cultural heritage organizations
  - Institutional Repositories
  - Government documents

- Integrate with library tools and systems
  - Connexion
  - WorldCat.org
  - WorldCat Local
Digital Repositories
cooperative advantage

- Integrate with other content
  - Journals
  - Books, manuscripts
  - Audio and video

- Increase the value, reduce the costs
  - Search across institutional collections
  - Search within the larger scope of content
    - WorldCat.org
  - Reduce cost through shared services
Future Direction
- New OCLC Strategy

Enterprise product architecture
Future direction - Products and services to support the collections grid

Collections Grid
A framework for representing content

Published Content
- Books
- Journals
- Newspapers
- Gov. docs
- CD, DVD
- Maps
- Scores

Special Collections
- Rare books
- Local/Historical newspapers
- Local history materials
- Photographs
- Archives & Manuscripts
- Theses & Dissertations

Open Web Content
- Freely-accessible web resources
- Open source software
- Newsgroup archives
- Images

Institutional Content
- ePrints/tech reports
- Learning objects
- Courseware
- Local government reports
- Training manuals
- Research data

Source: OCLC Office of Research, 2003
Future direction - support for the full range of library content

- dContent
- eContent
- “Digital Stacks”
Thank you

Questions?

email:Greg_Zick@oclc.org