Using DSpace for Digitized Collections

Lisa Spiro, Marie Wise, Sidney Byrd & Geneva Henry
Rice University

Open Repositories 2007
January 23, 2007
Outline of presentation

- Rationale for choosing DSpace
- Rice digitization projects using DSpace
- Obstacles to using DSpace for digitization projects
- Getting digitized content into DSpace
- Using hierarchies in a non-hierarchical system
- XML text support
- “Skinning” with Manakin
- Future plans:
  - Streaming media
  - JPEG 2000
Rationale for using DSpace

- Active open source community
- Support for digital preservation
- Single platform for all scholarly assets, both born-digital and digital
- Scalability
- Leverage efforts to integrate Sakai (for teaching assets) with DSpace (for longer-term institutional assets)
Digitization Projects in DSpace

- Travelers in the Middle East Archive (TIMEA): collection of XML-encoded texts, images, and maps focused on Western interactions with the Middle East. Supported by IMLS and CITI.
- Shepherd School Archive: digital audio of performances at the music school
- Rice Institute Pamphlets Archive: PDFs and XML-encoded text of a significant academic journal
- William Ward Watkin Architecture Collection: images & documents focused on Rice’s first architect
- Americas Archive: images & texts documenting hemispheric relations
- Rice Thresher Archive: back issues of the student newspaper
- Museum of Houston: artifacts from multiple repositories documenting Houston’s history
- Digitized Theses: existing theses from Rice’s founding to 1996
Obstacles to using DSpace for digitization projects

- DSpace was originally designed for born-digital, simple items (like PDFs).
- Lacks support for deeply hierarchical collections
- Only supports as-is content presentation
  - Transformation to HTML: TEI-encoded XML texts
  - Enhanced features: JPEG 2000, streaming media
- Needs customizable UI (Manakin)
- Integrating with other repositories/services
  - As part of larger collection (TIMEA)
  - Directly to other services (Yahoo Maps)
Getting digitized content into DSpace

- Digital objects created (scanning, etc.)
- Metadata created
  - Excel spreadsheet
  - Exported from legacy content management system
  - Created by vendor doing digitizing
- Metadata and file structure converted to DSpace batch import format
  - Import script normalizes dates, reformulates metadata as needed (splits fields, transforms to DC), matches metadata to objects, separates thumbnails, generates reports
- Admin web interface used to set up root Collection
Batch Importer

- Batch import command processes the files into DSpace
- Enhancements to import command:
  - File descriptions
  - Primary Bitstreams
  - Creating Communities and Collections
  - Testing that metadata fields are valid
  - Future: specify access permissions
- Web services ingestion would also be useful
Using hierarchical content in DSpace

- Can represent structured content in metadata or in DSpace Communities and Collections
- DSpace’s hierarchies (Community, Collection) are based on organizational units, not content
  - DSpace hierarchies not sufficiently flexible for representing structured content from some collections
    - Collections only hold Items
    - Communities only hold Collections and other Communities
    - Cannot mix Items and containers at same level
Adapting vocabulary and structures for DSpace

- **The Rice Institute Pamphlet**
  - Index to Rice Institute Pamphlet, Vols. 1-46

- **Volume 01**
  - Index to Volume One
  - Volume One, No. 1
  - Volume One, No. 2
  - Volume One, No. 3
  - Volume One, No. 4

- **Volume 02**
  - Index to Volume Two
  - Volume Two, No. 1
  - Volume Two, No. 2
  - Volume Two, No. 3
  - Volume Two, No. 4

- **Volume 03**
  - Index to Volume Three
  - Volume Three, No. 1
  - Volume Three, No. 2
  - Volume Three, No. 3
  - Volume Three, No. 4

- **Volume 04**
  - Index to Volume Four
  - Volume Four, No. 1
  - Volume Four, No. 2
  - Volume Four, No. 3
  - Volume Four, No. 4

Community: all RIP materials
Sub-community: each volume
  - Collection: volume index
    - Single index PDF for all four issues
  - Collection: each issue
    - Items for each article
Watkin Collection
Structure

Sub-communities within this community

- **Series I. Biographical, 1903-1953.**
  Forms part of the Rice Institute Pamphlets

- **Series II. Architectural Career, 1910-1952.**
  Forms part of the Rice Institute Pamphlets

- **Series III. Academic Career, 1914-1949.**
  Forms part of the Rice Institute Pamphlets

- **William Ward Watkin materials from Houston Metropolitan**
  Forms part of the Rice Institute Pamphlets

- **Wilmer Waldo & the Rice Institute Tunnel System**
  Forms part of the Rice Institute Pamphlets

Collections in this community

- **Correspondence**
  Forms part of the Rice Institute Pamphlets

- **Family**
  Forms part of the Rice Institute Pamphlets

- **Personal**
  Forms part of the Rice Institute Pamphlets
Linking to related materials

- Need to associate digital object with related materials, such as:
  - Book from which image is taken
  - Related teaching modules
  - Other editions of same book

- Currently accomplished via relation field in Dublin Core record
### Example

- TIMEA stylesheet in Manakin presents certain `dc.relation` fields as HTML links

<table>
<thead>
<tr>
<th>Standard interface:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>dc.relation.isreferencedby</strong></td>
</tr>
<tr>
<td><strong>dc.relation.isreferencedby</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>with Manakin stylesheet:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Related Resource:</strong></td>
</tr>
<tr>
<td><strong>Related Resource:</strong></td>
</tr>
<tr>
<td><strong>Related Resource:</strong></td>
</tr>
</tbody>
</table>
Future solutions: METS Viewer?

- DSpace supports storing METS data with each Item.
- METS would allow more nimble presentation of complex digitized objects such as:
  - Photo and postcard albums
  - Musical performances with multiple movements
  - Relationship between books and images within those books
  - Example: RLG METS viewer
- Need easier techniques for creating METS data
  - Inter-item METS references are harder because items must be referenced by handle — chicken and egg problem
- Work underway (China Digital Museum Project, other initiatives)
RLG METS Viewer

Ch'ên Hung-shou
Untitled (album of birds, flowers, and landscapes)
Ming Dynasty
album: ink and colors on silk
Kunstliebe St. Museum, Pacific Film Archive, University of California, Berkeley, 2026 Bancroft Way, Berkeley, California, 94720 United States

Created
Date: Ming Dynasty
People: Ch'ên Hung-shou
Places: China

Work
Untitled (album of birds, flowers, and landscapes)
Gift of James Cahill
Material and Technique: album: ink and colors on silk
Dimensions: w 6 x h 8 1/4 inches
Painting: paintings

Related Topics
bird and flower paintings
(Subject)

Collection
Ch'ên Hung-shou - Untitled album of birds, flowers, and landscapes
album leaf 1
album leaf 2

Email the image

<< First  Prev  Next >>  Last >>
XML support

- DSpace treats XML as “supported” format, but raw XML isn’t appropriate for presentation to users.
- Need to transform XML (TEI Light) to HTML for display in web browser
- Could store pre-generated HTML, but the HTML is just for presentation, not archival
XML vs HTML reading version

This XML file does not appear to have any style information associated with it. The document tree is shown below:

```xml
<TEI 2 TEIForm="TEI 2">
  <teiHeader TEIForm="teiHeader" status="new" type="text">
    <fileDesc TEIForm="fileDesc">
      <titleStmt TEIForm="titleStmt">
        <title TEIForm="title">
          Egypt through the stereoscope: a journey through the land of the Phara
        </title>
        <author TEIForm="author">Breasted, James Henry, 1865-1935</author>
      </titleStmt>
      <respStmt TEIForm="respStmt">
        <resp TEIForm="resp">Creation of machine-readable version</resp>
        <name TEIForm="name">TechBooks</name>
        <resp TEIForm="resp">Creation of digital images</resp>
        <name TEIForm="name">Electronic Resources Center, Fondren Library, Rice University</name>
      </respStmt>
      <respStmt TEIForm="respStmt">
        <resp TEIForm="resp">Conversion to TEI 2-conformant markup</resp>
        <name TEIForm="name">TechBooks</name>
        <resp TEIForm="resp">Parsing and proofing</resp>
        <name TEIForm="name">Electronic Resources Center, Fondren Library, Rice University</name>
      </respStmt>
      <respStmt TEIForm="respStmt">
        <resp TEIForm="resp">Distributor</resp>
        <name TEIForm="name">Rice University</name>
      </respStmt>
    </fileDesc>
  </teiHeader>
</TEI 2>
```

which you see scattered through the cemetery: they are erected over the tombs of the departed members of the family. Others who have no such dwelling erect a tent over the tomb. On at least one of the three days of the celebration, they come to the cemetery laden with palm branches which they strewn upon the tombs. The house just before us shows how such dwellings are arranged. The court covered over with plaited branches contains the tombs of the dead, while the roofed portion is the dwelling in which the relatives stay during the sojourn at the cemetery.

The hum of voices reciting the Koran, the shouts and the gay laughter and the rejoicing of the poor at the gate of the cemetery, as they receive the food distributed by the rich, at this, with the citadel and its splendid mosque outlined against the bluest of skies in the distance, gives the traveler a typical scene of oriental life, as it is found only at Cairo. Behind us at the Bab en-Nasr, which is also on this side of the city, the jubilation and merry-making are even more marked than here, and the temporary booths, with piles of sweets, the merry-go-rounds, the dancers and the rejoicing multitudes give are the impression of a large country fair.

In a few days these same people will be following the pieces of the sacred carpet or "kiswa" from the citadel to the Mosque of Hasanîn, in a rejoicing procession to which all Cairo will turn out. That procession is one of the most interesting public events at Cairo, and we shall later have the opportunity of observing it.
Each item only contains the TEI-encoded XML and the images.

Each TEI Item specifies which XSLT stylesheet is needed to transform it to HTML.
  - Each stylesheet is stored with the server config.
  - Can change one stylesheet in the config to update presentation of all dependent TEI Items

DSpace caches transformed HTML in special Bundle until Item or stylesheet changes
Text + Images = Sluggishness

- Researchers wanted original page thumbnails (for 300 + pp books) in HTML output.
- Serving thumbnails from DSpace slowed the server to crawl.
- Translating <img> links to use DSpace “retrieve” URLs helps.
- Even faster: translate <img> links to point to copies of thumbnails served directly by Apache.
  - This is acceptable since the HTML version is just presentation, not archival content.
Using DSpace as part of portal system

TIMEA includes:

- Texts, images & historical maps stored in DSpace
- GIS maps made available through ESRI’s ArcIMS
- Research & teaching modules published through Connexions
TIMEA portal

- Harvests OAI from DSpace and Connexions using OCLC Harvester 2.0
- NIMA Geonames used for GIS map
- Search of metadata & full text powered by Lucene
- Use Manakin to produce TIMEA UI for DSpace records
TIMEA portal
“Skinning” with Manakin

- “Brand” each collection with unique look and feel
- For TIMEA (timea.rice.edu):
  - Portal integrating different systems required TIMEA specific navigation being built into item-level records
  - Usability testing indicated that users had difficulty navigating if the item results lacked the same look & navigational elements
- Also re-themed GIS and Connexions to match TIMEA look
Hooray for Manakin!

- Developed by Texas A&M’s Digital Initiatives team
- Easy to create custom designs and interfaces
- Helps with structural problems (e.g. render “Collection” as “Issue”)
- Allows access to an XML representation of each page (add “/DRI” to the URL)
  - Provides information about individual files and their place in the Item’s structure
  - Sometimes preferable to OAI-PMH for exposing DSpace content to other systems.
TIMEA: Standard UI vs. Manakinized
Future plans for enhanced delivery of DSpace content

- Shepherd School Archive: digitizing magnetic reel tape
  - Only have streaming rights
  - Have DSpace generate reference file pointing browser to video/audio stream
  - Streaming server shares storage with DSpace

- Many large images of maps, posters
  - Better viewable if users can zoom and manipulate images in browser
  - Storing archival TIFF and also smaller copy for viewing is unnecessary duplication
  - JPEG 2000 software - solves both problems
Final thoughts

- Need to represent structure of content in DSpace
- Manakin transforms interface; need transformations for content
Acknowledgments

- IMLS and CITI for support of TIMEA
- Fondren Library staff
  - Brian Surratt
  - Amanda Focke
  - Debra Bailey Kolah
Contact Information

- Lisa Spiro: lspiro@rice.edu
- Sid Byrd: sidb@rice.edu