Fedora Commons or DSpace: A Comparison for Institutional Digital Content Repositories

Presented by: Aubrey Maynard, Laura Gentry, Adam Mosseri, Courtney Whitmore, Margaret Diaz, Camille Chidsey, and Kelly Kietur

WSU SLIS, WSU Library System, Technical Resource Center, and WSU NDSA Collaborative Project

Executive Summary

Challenges: Money, Manpower, Skills and Knowledge, Time

Overview: Fedora project and Dspace project

Challenges

Overview – Fedora Commons

- Project Details: Digitization and Ingest of the Detroit Sunday Journal, a weekly paper published by striking workers from The Detroit Free Press and The Detroit News from November 19, 1995 through November 21, 1999

STANDING STRONG FOR 1,592 DAYS — THE FIGHT'S NOT OVER YET

NOVEMBER 21-27, 1999 THE DETROIT VOL. 5 NO. 1 75 CENTS

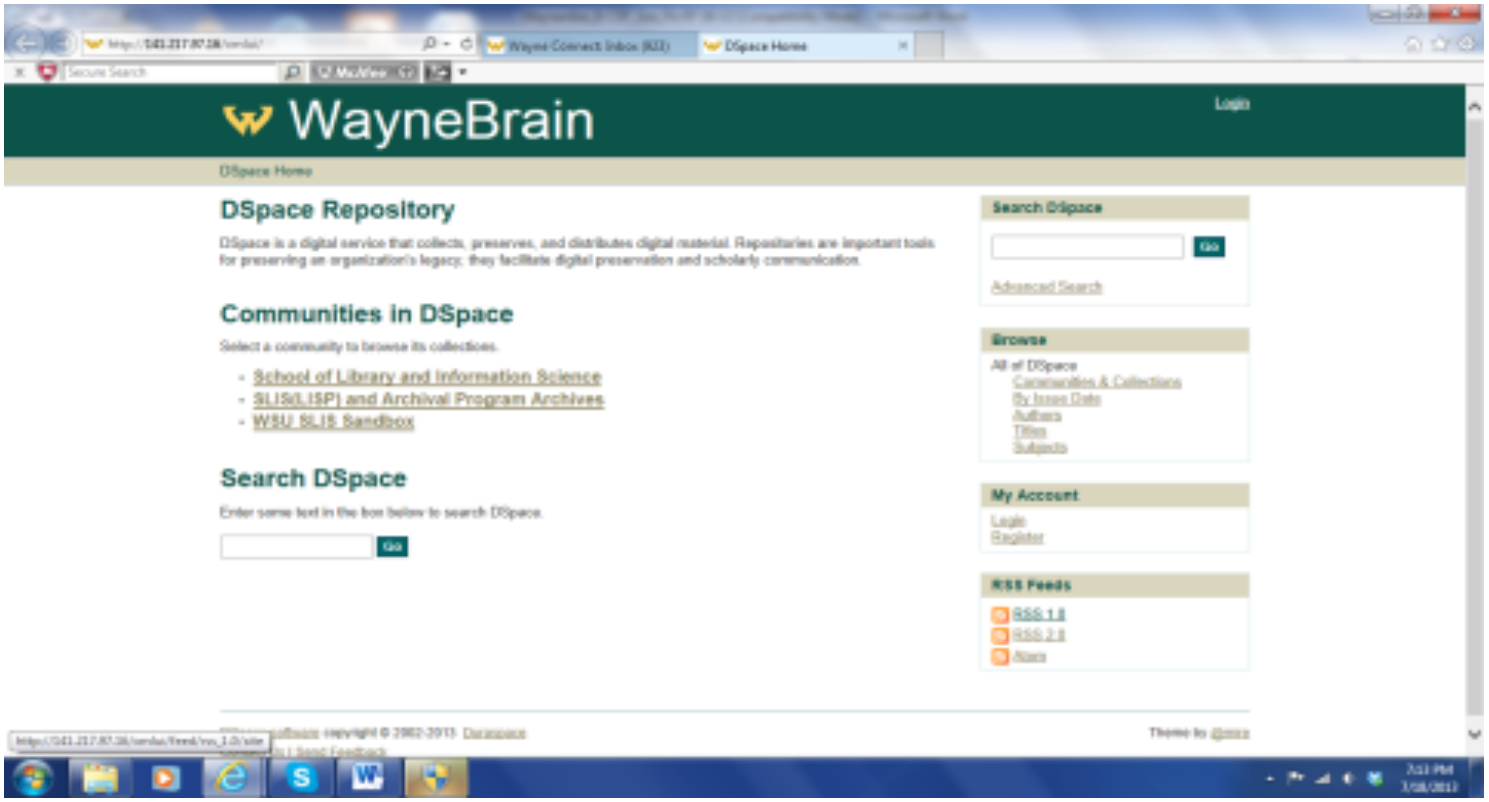
SUNDAY JOURNAL

PUBLISHED BY LOCKED-OUT DETROIT NEWSPAPER WORKERS

Our labor of love

Overview - DSpace

- Project Details: Digital archive of the School of Library and Information Science program at Wayne State University

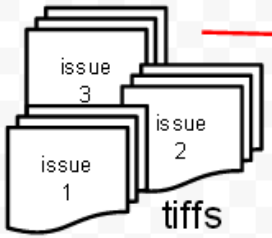


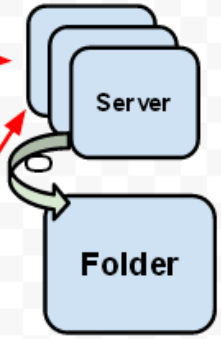
Fedora Commons: Overall Process and Workflow

- Digitalization scanning
- Creation of Metadata (MODS)
- Abbyy verification
- FOXML files
- Ingest in Fedora
- Ebook reader final product

Detroit Sunday Journal Workflow

1. Digitize





* Scans for an issue are housed in a folder, e.g... "***sunday_journal_v1iss5***"

The files are contained in that folder as they move around, scanning to "processing location".

3. Abbyy

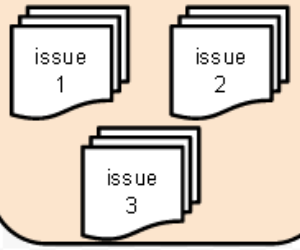
OCR & Text Verification

Abbyy output:
jpegs, html, xml, pdf, etc.
(much smaller than tiffs)

2. Metadata (MODS)

```
▼<mods:modsCollection
 xmlns:xlink="http://www.w3.org/1999/xlink"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-
 instance" xmlns:mods="http://www.loc.gov/mods/v3"
 xsi:schemaLocation="http://www.loc.gov/mods/v3
 http://www.loc.gov/standards/mods/v3/mods-3-4.xsd">
  ▼<mods:mods version="3.4">
    ▼<mods:titleInfo>
       <mods:nonSort>The</mods:nonSort>
       <mods:title>Detroit Sunday Journal:
       </mods:title>
```

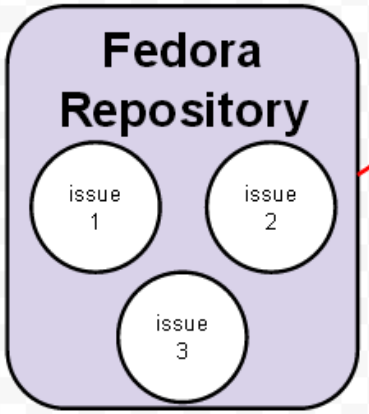
Processing Location Server



4. Fedora Object XML (FOXML)

```
▼<foxml:digitalObject
 xmlns:foxml="info:fedora/fedora-system:def/foxml#"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-
 instance"
 xmlns:database="http://www.oclc.org/pears/"
 xmlns:mods="http://www.loc.gov/mods/v3"
 xmlns:srw_dc="info:srw/schema/1/dc-schema"
 VERSION="1.1"
 xsi:schemaLocation="info:fedora/fedora-
 system:def/foxml#
 http://www.fedora.info/definitions/1/0/foxml1-1.xsd"
 PID="wayne:vol02no0119961117">
 ▼<foxml:objectProperties>
    <foxml:property NAME="info:fedora/fedora-
    system:def/model#state" VALUE="A"/>
    <foxml:property NAME="info:fedora/fedora-
    system:def/model#label"
    VALUE="vol02no01 19961117"/>
```

5. Ingest into Fedora Commons



6. E-book Reader

1. Digitize

- Both physical and digital holdings
- Scanned and saved as TIFFs
- Standardized naming conventions
  - Top-Level Folder:

    sunday_journal_v1iss5
  - Folder:

    vol#no#_YEARMMDD

    vol03no16_19980301
  - Scan:

    vol#no#_YEARMMDD_pg#.tif

    vol03no16_19980301_pg7.tif
- Uploaded and stored on a server

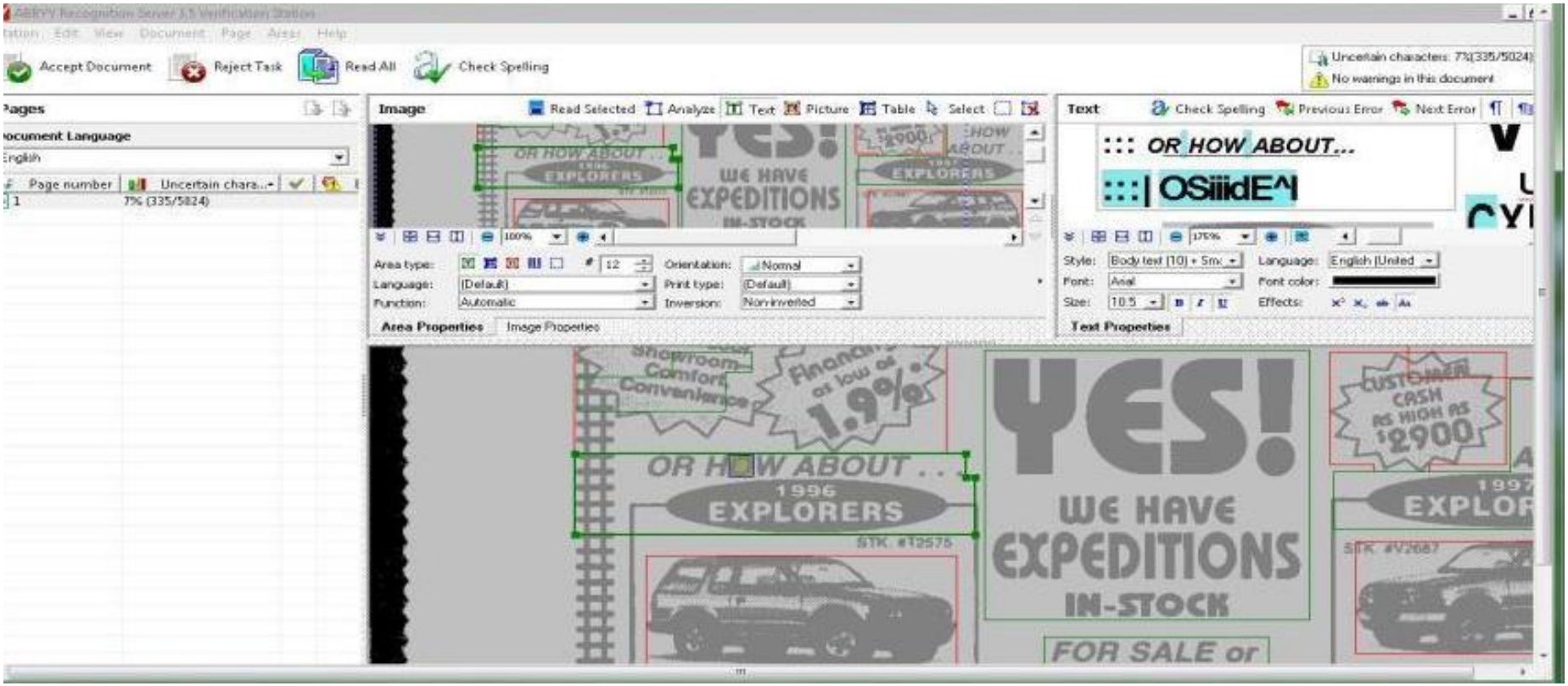
2. Metadata (MODS)

- FTP Client to download and upload from server
- Metadata Object Description Schema (MODS)
  - Written and saved in XML
  - Used a template
  - Validated

```
▼<mods:modsCollection xmlns:xlink="http://www.w3.org/1999/xlink"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns:mods="http://www.loc.gov/mods/v3"
 xsi:schemaLocation="http://www.loc.gov/mods/v3
 http://www.loc.gov/standards/mods/v3/mods-3-4.xsd">
  ▼<mods:mods version="3.4">
    ▼<mods:titleInfo>
       <mods:nonSort>The</mods:nonSort>
       <mods:title>Detroit Sunday Journal:</mods:title>
      ▼<!--
          Add the date of the issue, written out,
                                        transcribed from front page
                                        example: October 18 - 24, 1998
       -->
```

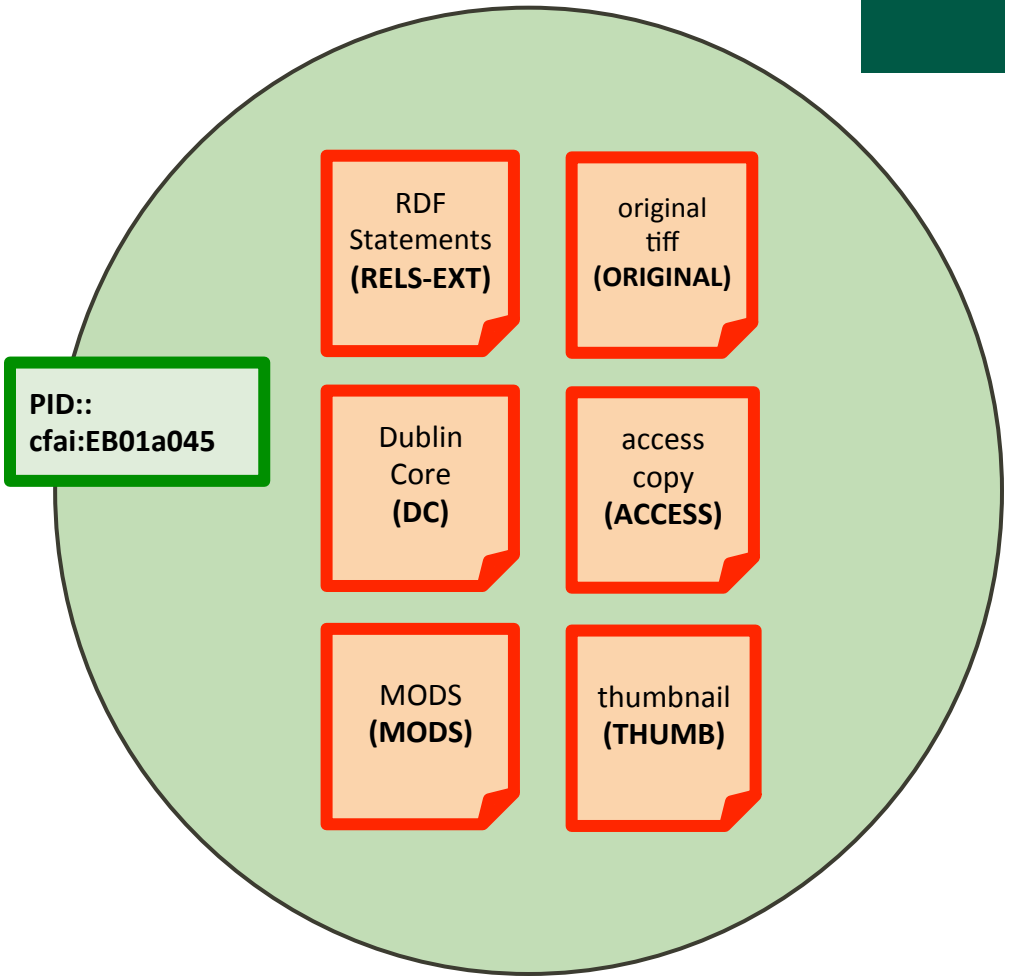
3. Abbyy

- OCR and Text Verification
- Below 5% Uncertainity
  - Car Ads, Error Messages, and No Jobs Left
- Output
  - Jpegs, html, xml, pdf, etc.
  - File size smaller than tiffs



4. Fedora Object XML (FOXML)

- Wrapper for XML and all the related metadata (RDF, Dublin Core, MODS, etc.)
- Uniting data streams into one long, complex file
- Object blueprint
- Need for ingest into Fedora Commons



Within FOXML: RDF Statements

- Describes relationships between objects
- RDF Statements also knows as triples (subject, predicate, and an object)

In computer speak...

subject = <fedora:ramsey:Sketchesandscraps>predicate = <fedora-rels-ext:isMemberOfCollection>object = <fedora:collection:ramsey>

In human speak...

The ebook object, "Sketches and Scraps", is a member of the Collection "Ramsey".

```
▼<rdf:Description rdf:about="info:fedora/wayne:vol01no0119951119">
   <fedora:isMemberOfCollection rdf:resource="info:fedora/wayne:collectionDSJ"/>
   <fedora:hasContentModel rdf:resource="info:fedora/singleObjectCM:WSUebook"/>
  </rdf:Description>
</rdf:RDF>
```

5. Ingest into Fedora Commons

- Use API (Application Programming Interface) to ingest in bulk
  - addDatastream
  - addRelationship
  - export
  - Ingest
- Giving instruction by command line on where to find everything
- Log into the Linux processing location server using SSH (Secure Shell) using Terminal or Putty
- STFP protocol from the command line
- Downloaded *Detroit Sunday Journal* issues from server

```
root@Silo:/var/www/wsuls/fedoraManager/uploads/FOXML# curl -i -u    username:password    -X POST http://localhost/fedora/objects/new
 -H 'Content-Type: text/xml' --data-binary '@cfai_EB02a046.xml'
HTTP/1.1 100 Continue

HTTP/1.1 201 Created
Date: Thu, 18 Apr 2013 14:21:31 GMT
Set-Cookie: JSESSIONID=DB30438F82E4A63D7470ED47F3EB50A8; Path=/fedora/; HttpOnly
Location: http://localhost/fedora/objects/wayne%3ACFAIEB02a046
Content-Type: application/octet-stream
Content-Length: 18

wayne:CFAIEB02a046


root@Silo:/var/www/wsuls/fedoraManager/uploads/FOXML#
```

6. E-book Reader

- Success!
- Simple temporary front end for in-house use only
- Designed as a web interface to Fedora's ingest
  - Final web interface will also include features such as cross volume/ issue searchability and browsing by volume
- If ingest fails, troubleshoot what went wrong and try again

+ Fedora Commons: Installation

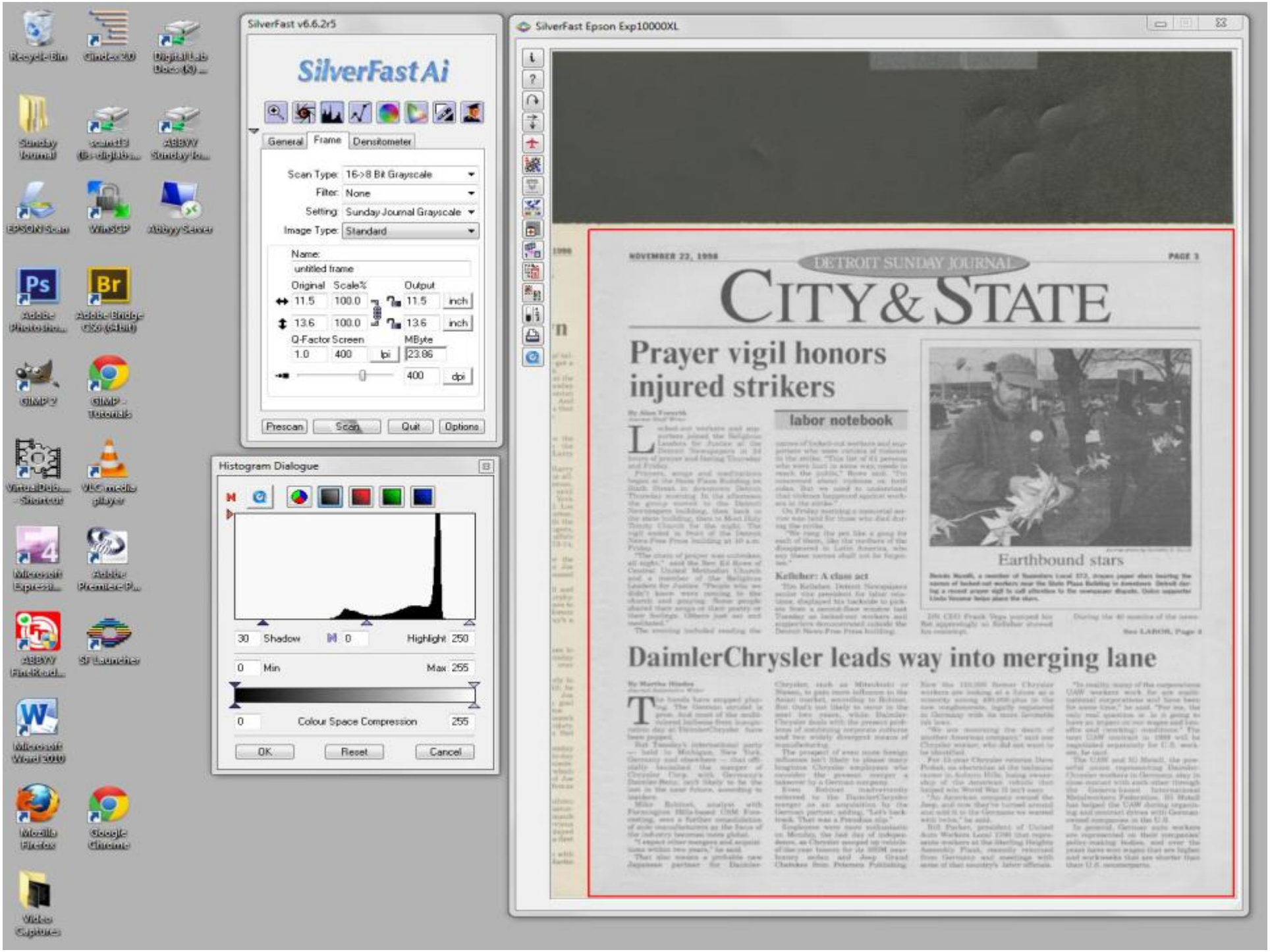
Technology Requirements
Skill Requirements

Basic Skills

- Skills
  - Basic digitization knowledge and understanding of standard scanning software and equipment.
  - How to manipulate and modify XML files through the use of an XML editor.
  - Competency reading and writing MODS files.
  - Basic understanding of computer networking, multiple user files, networked servers, and remote desktop interfaces.
  - Knowledge of OCR software and file ingests.
  - Good communication skills and people who work well together.

Technology Requirements

- Scanner with DPI capabilities of 300 or higher.
- Oversized scanner for large prints.
- Windows Vista O.S. or higher, MAC 10.1, or Linux based platformed computers.
- SSH editors (e.g. PuTTY or Terminal)
- XML editors (e.g. Notepad++, jEdit, Bluefish)
- Internet Connection 6mps or higher. 3mps or lower causes remote desktop controls to miscommunicate with the primary terminal causing delays in user processing capabilities.


SilverFast v6.6.2r5
SilverFastAi
General
Frame
Densitometer
Scan Type: 16->8 Bit Grayscale
Filter: None
Setting: Sunday Journal Grayscale
Image Type: Standard
Name:
untitled frame
Original
Scale%
Output
Q-Factor Screen
MByte
Prescan
Scan
Quit
Options
Histogram Dialogue
Shadow
Highlight
Min
Max
Colour Space Compression
OK
Reset
Cancel
SilverFast Epson Exp10000XL
NOVEMBER 22, 1998
DETROIT SUNDAY JOURNAL
PAGE 3
CITY & STATE
Prayer vigil honors injured strikers
labor notebook
Earthbound stars
DaimlerChrysler leads way into merging lane

Digitization

- Process
  - Scan each journal individually page by page.
  - To decrease the size of the .Tiff files and to best utilize our storage capabilities, it was determined that only pages with color (e.g. The Covers) were to be digitized in color formatting and subsequent pages were to be in black and white.
  - The average black and white file size was around 16,000 KB and a color file size was on average 60,000 KB.
  - After the entire journal was fully digitized it would be uploaded onto the networked drive and then the journal would be checked off on a master list indicating its completion.

OCR Abbyy

- Set Terminal Desktop Computer
- Coordination
- VPN Connection
- Open Verification Station
- 5% Uncertainty Limit
- Accept the document and mark as complete on master list
- Exit out of remote desktop connection

Fedora Commons: Metadata Issues

Why MODS and XML?

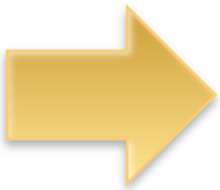
- In Fedora, documents are represented by an XML file; MODS is an XML schema
- Metadata in MODS:
  - Is comparatively end user friendly
  - Can represent MARC records simply
  - May be used in circumstances where the metadata will be packaged with an electronic resource
  - Disadvantages: Cannot readily be converted back to MARC record without loss of specificity , and sometimes without loss of data

Subject Headings

- Translating LCSH into MODS

LC control no.: n 50031280
LCCN permalink: http://lccn.loc.gov/n50031280
HEADING: Hoffa, James R. (James Riddle), 1913-

000 00557cz a2200181n 450
001 3264515
005 19870506125813.4
008 800625n| acannaab |a aaa
010 __ |a n 50031280
035 __ |a (DLC)n 50031280
040 __ |a DLC |c DLC |d DLC
100 10 |a Hoffa, James R. |q (James Riddle), |d 1913-
400 10 |w nna |a Hoffa, James Riddle, |d 1913-
400 10 |a Hoffa, Jimmy, |d 1913-
670 __ |a Martin, J. B. |b Jimmy Hoffa's hot, 1959.
670 __ |a Franco, J. Hoffa's man, c1987: |b CIP t.p. (in other info.: James R. Hoffa)
952 __ |a RETRO
953 __ |a xx00 |b ba30



```
48 <mods:subject authority="lcsh">
49 <mods:name authority="lcsh" type="personal">
50 <mods:namePart>Hoffa, James R. (James Riddle)</mods:namePart>
51 <mods:namePart type="date">1913-</mods:namepart>
52 </mods:name>
53 </mods:subject>
```

For example…

World War, 1939-1945 ‡x Women ‡z United States (World War, 1939-1945 --- Women --- United States)

```
<mods:subject authority="lcsh">
        <mods:topic>World War, 1939-1945</mods:topic>
        <mods:topic>Women</mods:topic>
         <mods:geographic>United States</mods:geographic>
</mods:subject>
```

Dashes representing subdivisions become like a new “paragraph”, indicating a separate tag in MODS

Newspaper Strike, Detroit, Michigan, 1995-2000

LCSHs lacking dashed subdivisions, which are simply refined, are often left as is.

```
<mods:subject authority="lcsh">
        <mods:topic>Newspaper Strike, Detroit, Michigan, 1995-2000</mods:topic>
</mods:subject>
```

Subject Headings

- Deciding when and when not to use subject headings
- Refining subject headings
  - What “details” or subdivisions are necessary
  - Temporal tags
  - Simple headings are OKAY.
- The Undecided Blanket Subject Heading

```
53 </mods:subject>
54 <mods:subject authority="lcsh">
55 <mods:geographic>Detroit (Mich.)</mods:geographic>
56 <mods:topic>Newspapers</mods:topic>
57 </mods:subject>
```

“Special Issue” Issues

Brought up discussion of:

- Alteration of the template
- Change in naming conventions
- Appropriate subject headings

2.30.	6/8/1997
2.31	6/15/1997
2.31SI	6/15/1997
2.32	6/22/1997
2.33	6/29/1997

```
- <mods:detail type="issue">
      <!-- If the special issue doesn't include an issue
      number put [special issue] into the number field ex:
      <mods:number>[special issue]</mods:number> -->
    <mods:number>[special issue]</mods:number>
    <mods:caption>no.</mods:caption>
```

```
<mods:subject authority="lcsh">
    <mods:topic>Detroit Red Wings (Hockey team)
        </mods:topic>
</mods:subject>
<mods:subject authority="lcsh">
    <mods:topic>Stanley Cup (Hockey)</mods:topic>
</mods:subject>
<mods:subject authority="lcsh">
    <mods:geographic>Detroit (Mich.)</mods:geographic>
    <mods:topic>Newspapers</mods:topic>
</mods:subject>
    <!-- If the special issue is about a particular event, add This
    special issue provides coverage of... to the end of the
    abstract -->
<mods:abstract> Issue of The Detroit Sunday Journal, a
    weekly newspaper published by striking workers
    during the Detroit Newspaper Strike. This special issue
    provides coverage of the Detroit Red Wings and their
    Stanley Cup victory. </mods:abstract>
```

Additional Issues with Naming Conventions

- Discrepancies in the Content Audit Document
  - Date discrepancies in the shared Google Spreadsheet document
  - Ripple affect to several parts of the project and work groups: scanning, metadata, errors that needed to be corrected on the server…

	Vol/Issue	Date	Scanned?	Own In Print	MODS created?	Notes- SJ; files; Tiff	Sent to Abbyy	Text Verified
1	3.42	8/30/1998	X	X	LMG			
2	3.43	9/6/1998	X	X	LMG			
3	3.44	9/13/1998	X	X	LMG			
4	3.45	9/20/1998	X	X	LMG			
5	3.46	9/27/1998	X	X	LMG			
6	3.47	10/4/1998	X		LMG			
7	3.48	10/11/1998	X		LMG			
8	3.49	10/18/1998	XX		LMG			
9	3.5	10/25/1998	XX		LMG			
0	3.51.	11/1/1998	X		LMG			
1	3.52	11/8/1998	X		LMG	Extra front page in B&W wi		
2	4.1	11/15/1998	X					
3	4.1	11/15/1998	KDK	X	AEM			

Welcome to **WayneBrain**

Installing and Using DSpace

WSU SLIS, WSU Library System, and
WSU NDSA Collaborative Project

DStarting Point

Background of the project's beginnings, statement of problem/issue we're exploring.

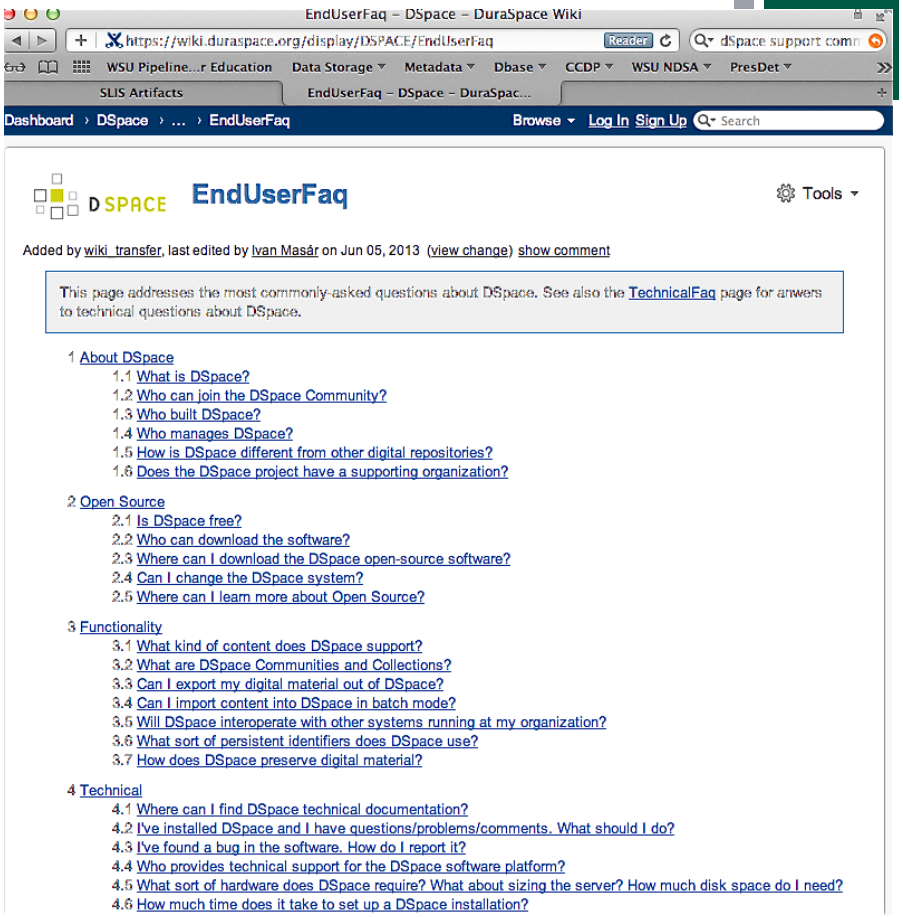
The Facts

- A “turnkey” digital repository
- Focused on long term storage, access and preservation of digital content."
- An “out of the box” solution to the problem of how/where to store digital materials.
- A place to store digital artifacts for later retrieval

WSU libraries on upswing; apply for reaccreditation with ALA

Why did SLIS choose DSpace?

- An addition to our Digital Projects Lab
- An opportunity for students to gain experience with digital repositories
- Free tech support!
- Fast installation
- Made a sandbox version in which to make mistakes



DProcess

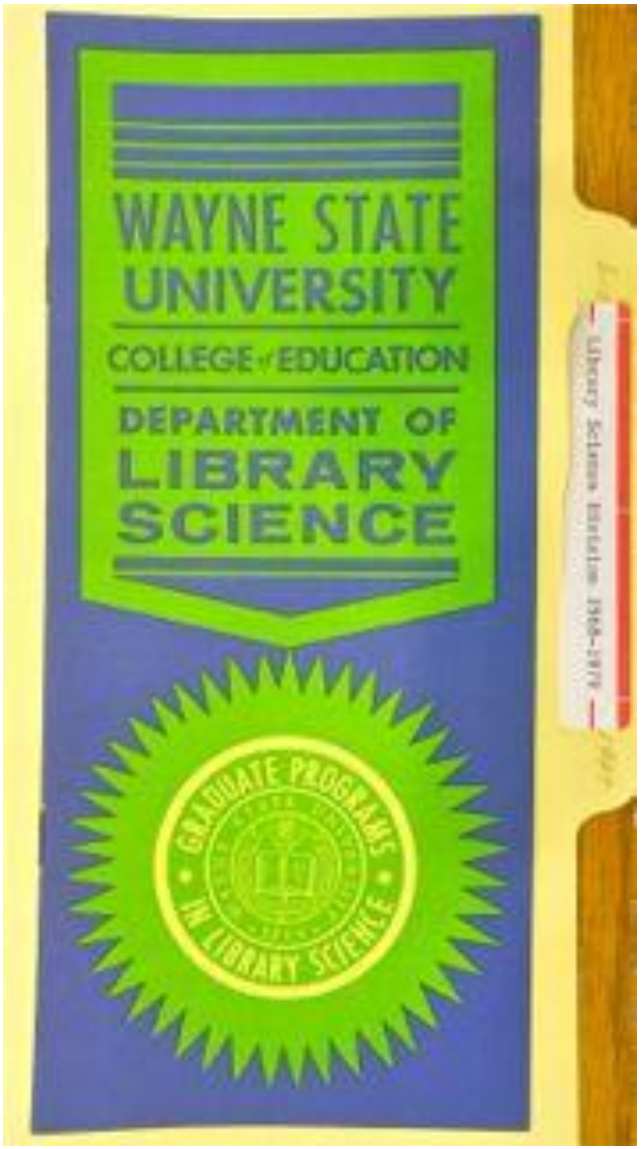
Review Install. Outline workflow, skills needed.

Installation

- Operating Systems independent (Windows or Linux)
- We use Linux
- Written in Java
- Stores information in PostgreSQL (relational database)
- Adheres to Dublin Core metadata standards
- XML user interface
- Done primarily by Bradley Woodruff, WSU Technology Assistant and WSU NDSA member

Basic Skills

- Skills
  - Basic digitization knowledge and understanding of standard scanning softwares and equipment.
  - Archival education helpful in selecting content.
  - Understanding of tagging and metadata.
  - Good communication and collaboration skills.
  - Patience!



1. Research & Content Selection

- Identifying collections and sub-collections for inclusion
- Exploring several different resources

  Direct student and alumni contact
- Identifying important faculty members
- Selecting materials for inclusion
- Format selection
- Ongoing process

2. Digitize

- Digitized physical holdings
  - Initial faculty publications were already digitized
- Scanned and saved as JPEG's
- Standardized naming conventions
- Simple numbering system
  - Year of digitization and then a sequential number
  - ex. 2013_00250
- Numbering system became accession #'s in DSpace

3. Metadata (Modified Dublin Core)

- DSpace configured with pre-existing DC fields
- Modified and adapted fields to fit our needs

dc.contributor.author	Turrini, Joseph M. (SLIS)	
dc.date.accessioned	2013-07-13T18:18:31Z	
dc.date.available	2013-07-13T18:18:31Z	
dc.date.copyright	2012	
dc.date.issued	2012	
dc.identifier.citation	Turrini, J. M. (2012). From History to Library and Information Science: A Case Study of Archival Education at Wayne State University. Information & Culture, 47(3), 358-380.	en_US
dc.identifier.uri	http://141.217.97.16/xmlui/handle/sliswsu/50	
dc.description.abstract	An article about the history and development of archival education at Wayne State University.	en_US
dc.language.iso	en_US	en_US
dc.subject	Wayne State University	en_US
dc.subject	Archives	en_US
dc.subject	Education	en_US
dc.title	From History to Library and Information Science: A Case Study of Archival Education at Wayne State University	en_US
dc.identifier.accnum	2013_00505	
dc.description.notes	Originally downloaded from Project Muse.	en_US
dc.publication.title	Information & Culture	en_US
dc.identifier.doi	10.1315/lac.2012.0019	
dc.type	Article	en_US

dc.contributor.author	Turrini, Joseph M. (SLIS)	
dc.date.accessioned	2013-07-13T18:18:31Z	
dc.date.available	2013-07-13T18:18:31Z	
dc.date.copyright	2012	
dc.date.issued	2012	
dc.identifier.citation	Turrini, J. M. (2012). From History to Library and Information Science: A Case Study of Archival Education at Wayne State University. Information & Culture, 47(3), 358-380.	en_US
dc.identifier.uri	http://141.217.97.16/xmlui/handle/sliswsu/50	
dc.description.abstract	An article about the history and development of archival education at Wayne State University.	en_US
dc.language.iso	en_US	en_US
dc.subject	Wayne State University	en_US
dc.subject	Archives	en_US
dc.subject	Education	en_US
dc.title	From History to Library and Information Science: A Case Study of Archival Education at Wayne State University	en_US
dc.identifier.accnum	2013_00505	
dc.description.notes	Originally downloaded from Project Muse.	en_US
dc.publication.title	Information & Culture	en_US
dc.identifier.doi	10.1315/lac.2012.0019	
dc.type	Article	en_US

4. Upload into WayneBrain

Title:

Enter the main title of the item.

Authors:

Enter the names of the authors of this item below.

Add

Last name, *e.g. Smith*

First name(s) + "Jr", *e.g. Donald Jr*

Contributors:

Person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors.

Add

Publisher:

Enter the name of the publisher of the previously issued instance of this item.

Series/Report No.:

Enter the series and number assigned to this item by your community.

Add

Series Name

Report or paper No.

Subject Keywords:

Enter appropriate subject keywords or phrases below.

Add

Subject Categories

DSurprises, DTips, & DTricks

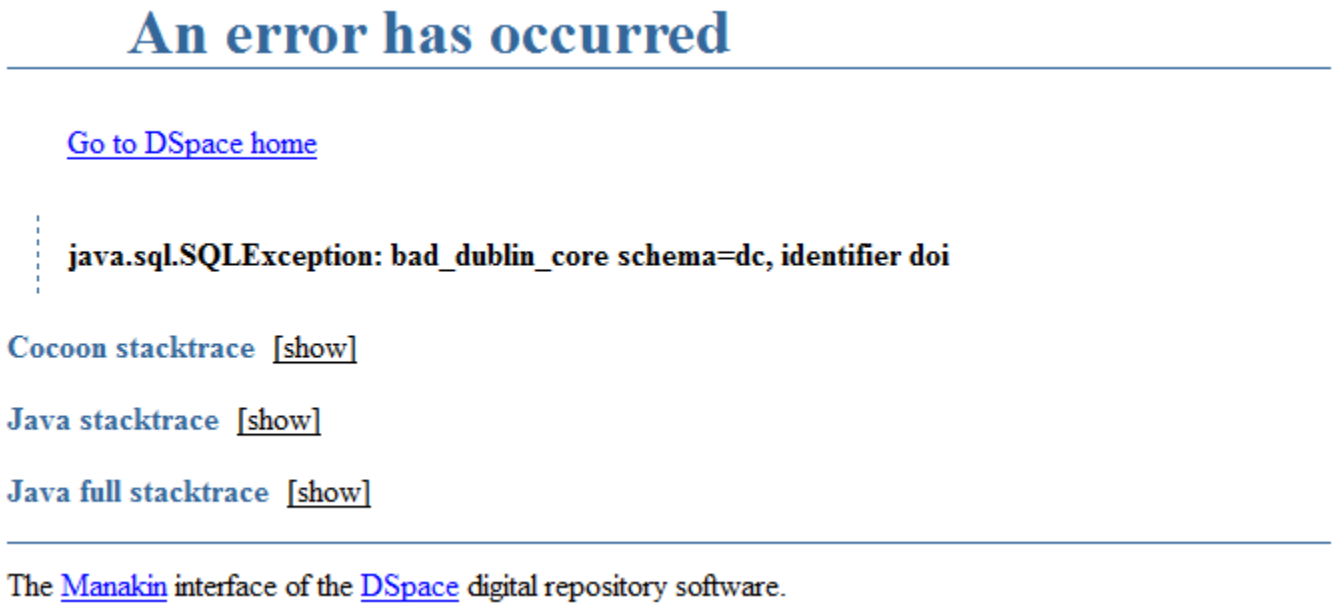
- Selecting Metadata Templates. Decisions, decisions….copyright (JSTOR example), DOI, handling related collections, search issues, controlled vocabularies for subjects, access, OCR for text, searchable PDF’s.

Subject Headings

- Refining subject headings
  - What “details” or subdivisions are necessary?
  - Temporal tags
  - Simple headings are OKAY.
- Make a controlled vocabulary!
- Pre-defined subject headings & tags are essential to continuation of project

Odds & Ends

- Uploading errors
- Admin rights



An error has occurred

Go to DSpace home

java.sql.SQLException: bad_dublin_core schema=dc, identifier doi

Cocoon stacktrace [show]

Java stacktrace [show]

Java full stacktrace [show]

The Manakin interface of the DSpace digital repository software.

Copyright and Licensing

- We needed to add a "rights" field in DSpace.
- Each individual artifact contains the "rights" statement in its metadata.
- WSU NDSA modeled their "rights" field off of the Smithsonian, another DSpace user.
- A longer version of their statement can be adapted for WSU's blanket licensing agreement.

DConclusions

What we like. Why pick Dspace? Who/ what is it suited to? Students' perspective on how it's different than Fedora. Tool for student training.

Why Choose DSpace?

Open Source. It's FREE!

Large user base (Over 1000 institutions currently use DSpace)

- Deep Blue (University of Michigan)

  Deepblue.lib.umich.edu

- The Smithsonian

  Si-pddr.si.edu/dspace

- Libraries, government agencies, etc

Handles multiple formats of digital content (images, documents, videos, programs, etc)

Contents are indexed by Google

Built-In Digital Preservation Functions

- Deep research community supporting DSpace in its digital preservation activities.
- Meets criteria for trusted repository
- Assigns checksums automatically

Student Perspectives

Wonderful student training tool!

Easy way for students to practice using content management systems for digital preservation

Higher learning curve for Fedora due to technical skills and the creation of multiple objects, such as images, OCR files, metadata, etc…

"Fedora might be better suited for a project/group with a stable membership base, as the amount of steps/learning involved could slow down the process if new people had to learn every time. DSpace would be more suited to a student project because other than the initial install it's pretty much there and ready to go and its format is more conducive to turnover." - Kelly Kietur, WSU NDSA Research Co-Chair

Why Fedora Commons?

"We tried/experimented with Fedora because we had a lot of digital items. These digital items were growing in number and getting more complex; and most of the out-of-the-box solutions similar to Fedora didn't seem to fit what we needed."

-Cole Hudson, WSU Librarian

- It is modular so you can customize it for a particular technology for a specific task.
- From a preservation standpoint it is more adaptable to the changing technologies and of the current time.
- More control over how tour assets are managed.

Summary Statement

- "Preservation is a critical function for any institutional repository, as organizations both large and small realize the need for built-in digital preservation tools to ensure access, storage, and management for the long-term. Both DSpace and Fedora fulfill that responsibility admirably. The choice of which better suits an institution depends on its resources, technology support, and desire to have a heavily customizable or out-of-the box solution to institutional repository and preservation needs.“

-Lisa Phillips, WSU SLIS Student

DCollaboration & Acknowledgements

Teamwork and Collaboration on Fedora and DSpace Projects

- Wayne State University Faculty/Staff:
  - Kim Schroeder – Faculty Advisor
  - Joshua Neds-Fox – Coordinator for Digital Publishing, Wayne State University Library
  - Amelia Mowry – Metadata and Discovery Services Librarian, Wayne State University Library
  - Cole Hudson – Digital Publishing Librarian, Wayne State University Library
  - Graham Hukill – Digital Publishing Librarian, Wayne State University Library

Teamwork and Collaboration on Fedora and DSpace Projects - Students

- Adam Mosseri – Student / Fedora
- Aubrey Maynard – NDSA Chapter Vice President / Student / Fedora
- Camille Chidsey – NDSA Chapter President / Student / DSpace
- Courtney Whitmore – NDSA Research Co-Chair / Student / Fedora
- Kelly Kietur – NDSA Research Co-Chair / Student / Fedora and DSpace
- Kevin Barton – NDSA Digital Liaison / Student
- Laura Gentry – NDSA Secretary / Student / Fedora
- Lisa Phillips – Student / DSpace
- Lura Smith - Student / DSpace
- Margaret Diaz - Student / Fedora

Collaboration Tools

- Email
- Adobe Connect Meetings
- Fedora Documentation and Resources for Sunday Journal Project
- Google Docs Spreadsheet/Tracker
- Digital Media Projects Lab (scanning)
- VPN and Remote Desktop

Want to Hear More?

wsustudentndsa.wordpress.com

facebook.com/wsustudentndsa

@wsundsa

OR...

We're available for hire!