Strategies and Tools for Digital Repository Selection and Migration

Todd Crocken  
*University of Houston*

Anne Washington  
*University of Houston*

Follow this and additional works at: https://digital.sandiego.edu/symposium


This Workshop is brought to you for free and open access by Digital USD. It has been accepted for inclusion in Digital Initiatives Symposium by an authorized administrator of Digital USD. For more information, please contact digital@sandiego.edu.
Strategies and Tools for Digital Repository Selection and Migration

Presenter 1 Title
Content Strategist for the Bridge2Hyku project

Presenter 2 Title
Metadata Services Coordinator

Session Type
Workshop

Abstract
As digital repositories evolve, so to do the needs of institutions who employ them. Increasingly institutions are faced with the daunting task of migrating content from one repository to another. But what strategies exist to help institutions identify suitable repositories and effectively and efficiently plan and execute a migration? This workshop aims to explore the issues and strategies of repository (re)selection and migration.

Participants will learn about the different phases of a migration process including: system evaluation and selection, migration planning, and implementation strategies and tools. Throughout the workshop, participants will actively explore these phases as they relate to their organizational context and come away with questions and next steps for planning for system selection and/or a migration at their own institution.

The workshop will be led by members of the Bridge2Hyku (B2H) Project, an Institute for Museum and Library Services (IMLS) National Leadership/Project Grant (LG-70-17-0217-17) initiative led by UH Libraries to support the creation of the B2H Toolkit- a suite of resources for migration planning and implementation.

Location
KIPJ Room D

This workshop is available at Digital USD: https://digital.sandiego.edu/symposium/2019/2019/8
STRATEGIES AND TOOLS FOR DIGITAL REPOSITORY SELECTION AND MIGRATION

TODD CROCKEN - ANNE WASHINGTON
DIGITAL INITIATIVES SYMPOSIUM 2019
Basics

● 2 Year IMLS grant project (lg-70-17-0217-17)
● Based at University of Houston with partner institutions: University of Victoria, University of Miami, and IUPUI.
● Advisors/Stakeholders: Stanford, Indiana, DPLA, Duraspace.
Intros

Todd Crocken (he/him/his)
- Content Strategist for Bridge2Hyku
- 10 years of experience doing content migration outside academia
- Loves Tortas

Anne Washington (she/her/hers)
- Metadata Strategist for Bridge2Hyku
- 3 years on UH digital library migration project, 6 years experience with Samvera technologies
- Loves pizza

You? (Name, Job, Favorite Food)
Schedule

● System Evaluation and Selection ~ 60 mins
  ○ Short break ~10
● Migration Strategy ~ 60 mins
  ○ Short break ~10
● Migration Resources ~ 30 mins
Pairs, Groups, or ...?

- Lets self-select groups of mixed skills.
Goals

- Define evaluation criteria of your new platform by creating use cases and user stories.
- Identify your migration’s possible complexities through planning a roadmap
- Identify resources and tools for migration
Why Migrate?

- Cultural?
  - 
  - 
  - 

- Technical?
  - 
  - 
  - 

- Economic?
  - 
  -
Why Migrate?

● Cultural?
  ○ New emphasis on digitized collections
  ○ Digital Humanities

● Technical?
  ○ Metadata wrangling
  ○ Open source

● Economic?
  ○ Fees vs Skills
  ○ AWS
Why Migrate? – Rationales for UHL

● CONTENTdm sunsetted its local version in 2017.
● We needed a more robust DAMS for new digital initiatives
  ○ Is more flexible, scalable, interoperable
  ○ Manages larger amount of data in a variety of formats
  ○ Accommodates creative workflows
  ○ Allows for configuration of additional functionalities
  ○ Supports and enables linked data
Approaches to System Selection and Evaluation
WATERFALL

Discover  Plan  Build  Review

START  FINISH

AGILE

Discover  Plan  Discover  Plan  Discover  Plan

Build  Review  Build  Review  Build  Review

START  ITERATION  ITERATION  FINISH

**“WATERFALL”**

**UH Evaluation & Selection Process**

- Conducted needs assessment
  - Stakeholder requirements
    - Internal focus groups
  - Identified future needs
    - Reviewed aspirational institutions
    - Reviewed scholarly literature
  - Gathered data
    - Identified materials and users
UH Evaluation & Selection Process

● Developed selection criteria based on needs
  ○ System Function
  ○ Content Management
  ○ Metadata
  ○ UI/Search

● Established evaluation methods

“WATERFALL”
UH Broad Evaluation Methodology

Scoring:
- Scores were determined by reviewing supporting documentation, marketing information, or talks with vendors
- 0 - Does not support criteria
- 1 - Supports criteria
- Scores were totaled up and top 2 systems were evaluated for the final round
<table>
<thead>
<tr>
<th>DAMS</th>
<th>Evaluation Score</th>
<th>Possible Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fedora</td>
<td>27</td>
<td>29</td>
</tr>
<tr>
<td>Fedora / Hydra</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>Fedora / Islandora</td>
<td>26</td>
<td>29</td>
</tr>
<tr>
<td>Collective Access</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>DSpace</td>
<td>24</td>
<td>29</td>
</tr>
<tr>
<td>CONTENTdm</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>Rosetta</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>Trinity (iBase)</td>
<td>19</td>
<td>29</td>
</tr>
<tr>
<td>Preservica</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>Luna Imaging</td>
<td>15</td>
<td>29</td>
</tr>
<tr>
<td>RODA</td>
<td>6</td>
<td>29</td>
</tr>
<tr>
<td>Invenio</td>
<td>5</td>
<td>29</td>
</tr>
</tbody>
</table>
UH Detailed Evaluation Criteria

- Drawn from the same sources
- Looked at specific features for DAMS
"WATERFALL"

UH Detailed Evaluation Criteria

- Criteria was divided into eight testing sections
  - System Environment and Testing
  - Administrative Access
  - Content Ingest and Management
  - Metadata
  - Content Access
  - Discoverability
  - Report and Inquiry Capabilities
  - System Support
UH Evaluation Criteria

<table>
<thead>
<tr>
<th>Broad (29)</th>
<th>Detailed (131)</th>
</tr>
</thead>
<tbody>
<tr>
<td>● System Function</td>
<td>● System Environment and Testing</td>
</tr>
<tr>
<td>● Content Management</td>
<td>● Administrative Access</td>
</tr>
<tr>
<td>● Metadata</td>
<td>● Content Ingest and Management</td>
</tr>
<tr>
<td>● UI/Search</td>
<td>● Metadata</td>
</tr>
<tr>
<td></td>
<td>● Content Access</td>
</tr>
<tr>
<td></td>
<td>● Discoverability</td>
</tr>
<tr>
<td></td>
<td>● Report and Inquiry Capabilities</td>
</tr>
<tr>
<td></td>
<td>● System Support</td>
</tr>
</tbody>
</table>
System Setup for Detailed Evaluation

- Virtual servers were set up for each system
- Latest stable versions were chosen (no betas) and installed
- All supporting server software was installed
- Additional software to support the evaluation testing was installed and set up
Gathering Collections for Testing

- Wide variety of item formats currently available
- Variety of item formats for future projects
- Large collections
- Included items with bad metadata, wrong formats, and corrupted data
Detailed Evaluation Methodology

Scoring:
- Each system was tested with the same set of collection items
- Ranged score 0 - 3 (0 failed - 3 fully supported)
- Yes / No criteria was scored 0 or 3
- System documentation was still used in some scoring

Process:
- Scores were tallied to show top ranked systems
- Systems summarized into advantages and disadvantages
- Recommendation was compiled and written into the final report
# Score of Top 2 DAMS from UHL Testing Using Detailed Evaluation Criteria

<table>
<thead>
<tr>
<th>Testing Sections</th>
<th>DSpace Score</th>
<th>Fedora Score</th>
<th>Possible Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>System Environment and Testing</td>
<td>21</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>Administrative Access</td>
<td>15</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>Content Ingest and Management</td>
<td>59</td>
<td>96</td>
<td>123</td>
</tr>
<tr>
<td>Metadata</td>
<td>32</td>
<td>43</td>
<td>51</td>
</tr>
<tr>
<td>Content Access</td>
<td>14</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Discoverability</td>
<td>46</td>
<td>84</td>
<td>114</td>
</tr>
<tr>
<td>Report and Inquiry Capabilities</td>
<td>6</td>
<td>15</td>
<td>21</td>
</tr>
<tr>
<td>System Support</td>
<td>12</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td><strong>TOTAL SCORE:</strong></td>
<td><strong>205</strong></td>
<td><strong>300</strong></td>
<td><strong>393</strong></td>
</tr>
</tbody>
</table>
# Fedora/Hydra - Advantages and Disadvantages

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open source</td>
<td>Steep learning curve</td>
</tr>
<tr>
<td>Large development community</td>
<td>Long setup time</td>
</tr>
<tr>
<td>Linked Data ready</td>
<td>Requires additional tools for discovery</td>
</tr>
<tr>
<td>Modular design through API</td>
<td>No standard model for multi-file objects</td>
</tr>
<tr>
<td>Scalable, sustainable, and extensible</td>
<td></td>
</tr>
<tr>
<td>Batch import / export of metadata</td>
<td></td>
</tr>
<tr>
<td>Handles any file format</td>
<td></td>
</tr>
</tbody>
</table>
## DSpace - Advantages and Disadvantages

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open source</td>
<td>Flat file and metadata structure</td>
</tr>
<tr>
<td>Easy installation / turnkey system</td>
<td>Limited reporting capabilities</td>
</tr>
<tr>
<td>Existing familiarity through TDL</td>
<td>Limited metadata features</td>
</tr>
<tr>
<td>User group / profile controls</td>
<td>Does not support Linked Data</td>
</tr>
<tr>
<td>Metadata quality module</td>
<td>Limited API</td>
</tr>
<tr>
<td>Batch import of objects</td>
<td>Not scalable or extensible</td>
</tr>
<tr>
<td></td>
<td>Poor user interface</td>
</tr>
</tbody>
</table>
“WATERFALL”

Where UHL ended.

- Created a Report
- Left the ETDs in DSpace
- Digital Library
  - Hyku (Hydra-in-a-Box)
  - Hyrax
Do we have 6-12 months in this workshop?
Do we have 6-12 months in this workshop? (no)
Agile 101

- Software development-focused
- Project management framework
- Cross-functional team
- Incremental, iterative work
## Manifesto for Agile Solution Development

<table>
<thead>
<tr>
<th>“AGILE”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Individuals &amp; interactions</strong></td>
</tr>
<tr>
<td><strong>Working solutions</strong></td>
</tr>
<tr>
<td><strong>Customer collaboration</strong></td>
</tr>
<tr>
<td><strong>Responding to change</strong></td>
</tr>
<tr>
<td><strong>over</strong></td>
</tr>
<tr>
<td><strong>over</strong></td>
</tr>
<tr>
<td><strong>over</strong></td>
</tr>
<tr>
<td><strong>over</strong></td>
</tr>
<tr>
<td><strong>Processes &amp; tools</strong></td>
</tr>
<tr>
<td><strong>Comprehensive documentation</strong></td>
</tr>
<tr>
<td><strong>Contract negotiation</strong></td>
</tr>
<tr>
<td><strong>Following a plan</strong></td>
</tr>
</tbody>
</table>
MoSCoW Prioritisation

- Must Have
- Should Have
- Could Have
- Won’t Have this time
Minimum Viable Product

- What can the stakeholders and developers agree is the feature set required to launch the platform?
- “Bells and whistles” vs the actual car
What’s not viable about your current system?

“AGILE”
Use Cases?

- What are your broader institutional need?
  - ETD / IR?
  - Cultural Heritage Objects?
  - Special Collections?

- What are your unique needs?
  - Lots of data sets?
  - Lots of video?
  - Starting from scratch?

Worksheet time
User Stories

● Testable
● Much smaller unit of work/development

Basic Rules
● Don’t get technical
● Keep it short
● Be clear and avoid using phrases “maybe”, “something like”, etc.
● Avoid functional requirements, i.e. “click a button”
● Avoid conjunctions AND/OR
“AGILE”

User Stories

- Construction

As a WHO I want WHAT so that WHY.
How the user stories can help after this workshop?

- User stories are still a good way to bring your needs to consortiums and SaaS providers.
- Begin to think about using project management frameworks
- Kickstart asking the more nuanced questions about what your institution really needs from a new platform
BREAK
Migration Planning Worksheet

- Fill out during break, or while we discuss migrations in second part of this workshop
- We’re here to answer any questions
- It is ok to only have a vague answer
Between Selection and Migration

- Shouldn’t be a “baton passing”
- Should be the same group that makes the selection
  - But who are those people?
Stakeholders?

- Department Heads
- “Business Development” people
- Special Collections?
- Institutional Publisher?
Team Building

- Subject Matter Experts?
  - Special Collections?
  - Scholarly Communications?
- IT? Dev?
- Metadata Librarians
Timeline

● How long is the time between your selection and implementation?
  ○ What does “implementation” mean in your context?
    ■ Ready to migrate
    OR
    ■ All migration done
  ○ Do you need to reevaluate anything?
● MVP!
So you’re migrating?
Next up

- Migration Planning
- Normalize Metadata, Modify Content
- Migrate!
- Verify
Migration Planning

- Contextual Considerations
- Digital Library Analysis
- Content Analysis
- Metadata Analysis
Migration Planning

● Contextual Considerations
  ○ Type
  ○ Size
  ○ Resources
Migration Planning

● Contextual Considerations
  ○ Audience
  ○ Administration
  ○ Dependencies
Migration Planning

● Digital Library Analysis
  ○ Systems
  ○ Size of collection(s)
  ○ Data model
Migration Planning

- Content Analysis
  - Work types
## Migration Planning

### Content Analysis

<table>
<thead>
<tr>
<th>Collection</th>
<th>Objects</th>
<th>Photographic</th>
<th>Document</th>
<th>Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galveston 1915 Hurricane Photographs</td>
<td>124</td>
<td>Single-sided</td>
<td>Single-P</td>
<td>Hierarchical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Double-sided</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Multi-P</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Audio + Document</td>
</tr>
<tr>
<td>Burdette Keeland Architectural Drawings and Photographs</td>
<td>75</td>
<td>Single-sided</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Double-sided</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
Migration Planning

● Content Analysis
  ○ File types
  ○ File locations
Migration Planning

● Metadata Analysis
  ○ Schema use
  ○ Field requirements
  ○ Controlled vocabularies
Migration Planning

- Metadata Analysis
  - Copyright
  - Data quality
    - Consistency
    - Standards
Discussion

- Migration Planning
  - Contextual Considerations
  - Digital Library Analysis
  - Content Analysis
  - Metadata Analysis
Review and revise content

- Normalize metadata
- File management
- Other?
Review and revise content

- **Normalize metadata**
  - Why
    - Very important component of migration
    - Stakeholder communication
  - What
    - Identify and prioritize issues
Review and revise content

- Normalize metadata
  - How
    - Manual
    - Programmatic
  - When
    - Before
    - In transit
    - After
Review and revise content

- **File management**
  - File locations
  - Filenaming
  - Preservation
Discussion

- Review and revise content
  - Normalize metadata
  - File management
  - Other?
Migrate Content

- Map Metadata
- Model Work Types
- Source Repository Export
- Target Repository Import
Migrate Content

- Map Metadata

<table>
<thead>
<tr>
<th>Your Repository</th>
<th>Target Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title (dc.title)</td>
<td>Title (dc.title)</td>
</tr>
<tr>
<td>Description (dc.description)</td>
<td>Abstract (dct.abstract)</td>
</tr>
<tr>
<td>Creator (&lt;name&gt;&lt;namePart&gt;, etc..)</td>
<td>Creator (dc.creator)</td>
</tr>
</tbody>
</table>
Migrate Content

- Map Metadata

<table>
<thead>
<tr>
<th>Your Repository</th>
<th>Target Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title (dc.title)</td>
<td>Title (dc.title)</td>
</tr>
<tr>
<td>Description (dc.description)</td>
<td>Abstract (dct.abstract)</td>
</tr>
<tr>
<td>Creator (&lt;name&gt;&lt;namePart&gt;, etc..)</td>
<td>Creator (dc.creator)</td>
</tr>
</tbody>
</table>
## Migrate Content

- **Map Metadata**

<table>
<thead>
<tr>
<th>Your Repository</th>
<th>Target Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title (dc.title)</td>
<td>Title (dc.title)</td>
</tr>
<tr>
<td>Description (dc.description)</td>
<td>Abstract (dct.abstract)</td>
</tr>
<tr>
<td>Creator (&lt;name&gt;&lt;namePart&gt;, etc..)</td>
<td>Creator (dc.creator)</td>
</tr>
</tbody>
</table>
Migrate Content

● Model Work Types
Migrate Content

- Source Repository Export
- Target Repository Import
Content Verification

● Approaches
● Measures
BREAK
Migration Resources

● B2H Toolkit
  ○ Website
  ○ CDM Bridge
  ○ HyBridge
● Academic Articles
● Communities
The Website

- Migration best practices
- Online resources
  - Blogposts about specific issues
  - Links to other useful sites
  - Migration framework
- Toolkit instructions and info
- End of Phase 1&2 Reports
CDM Bridge
- Desktop Electron-app
- Works with hosted CDM
- Metadata crosswalking
- Export csv and files

- Ruby Gem for Hyrax and Hyku
- Creates easy to use tab for bulk ingest
- Allows collection-specific ingest
- Seamlessly integrated with CDM Bridge’s output

- **Breaking Up With CONTENTdm: Why and How One Institution Took the Leap to Open Source** - Heather Gilbert and Tyler Mobley (April 2013)
- **Evaluating, Planning, and Completing a Successful Migration: A Case Study** - Elizabeth Chance (2018)
- **Are we still working on this? A meta-retrospective of a digital repository migration in the form of a classic Greek Tragedy (in extreme violation of Aristotelian Unity of Time)** - Steve Van Tuyl, Josh Gum, Margaret Mellinger, Gregorio Luis Ramirez, Brandon Straley, Ryan Wick, Hui Zhang (August 2018)
Communities

- User Groups
  - DSpace User groups
  - Fedora User groups
  - CONTENTdm User groups
- Slack Servers
  - Samvera
    - #hydra-migration
  - Code4lib
  - Fedora-Project
  - Metadata Support Group
Extras? Comments? Thoughts?
CONTRIBUTE

- Join CONTENTdm Migration IG on samvera slack: #hydra-migrations
- Have a migration story for a blogpost or want to help with development/documentation?
  Contact Todd Crocken (tcrocken@uh.edu) or Andy Weidner (ajweidner@uh.edu)
- Follow us on Twitter! @BRIDGE2Hyku
- Check out our github: https://github.com/Bridge2Hyku
THANKS!

Background from unsplash.com
Abstract: As digital repositories evolve, so do the needs of institutions who employ them. Increasingly, institutions are faced with the daunting task of migrating content from one repository to another. But what strategies exist to help institutions identify suitable repositories and effectively and efficiently plan and execute a migration? This workshop aims to explore the issues and strategies of repository (re)selection and migration.

Participants will learn about the different phases of a migration process including: system evaluation and selection, migration planning, and implementation strategies and tools. Throughout the workshop, participants will actively explore these phases as they relate to their organizational context and come away with questions and next steps for planning for system selection and/or a migration at their own institution.

The workshop will be led by members of the Bridge2Hyku (B2H) Project, an Institute for Museum and Library Services (IMLS) National Leadership/Project Grant (LG-70-17-0217-17) initiative led by UH Libraries to support the creation of the B2H Toolkit—a suite of resources for migration planning and implementation.