Repository Migration
Stories: A Shared Knowledge Approach to Lowering Barriers

Seth Shaw – Arizona State University
Kate Dohe – University of Maryland
Julia Corrin – Carnegie Mellon University
Arran Griffith – Fedora Program
Complete & Partial System Migrations

Seth Shaw, Digital Library Software Developer
Replacing system components, all or in part, with new components.

Retaining (potentially) (mostly):

1. Content: metadata, documents, images, audio/video
2. Business Rules: intentional system constraints; e.g. permissions, data standards, and workflows.
3. Fundamental User Experience: e.g. search and view item with metadata
Complete System Migrations: Moving house

System Component “Major” Upgrades: Remodeling the kitchen
Generalized Steps:
- Exporting content
- Metadata mapping and remediation
- Software localization
- Hardware infrastructure provisioning
- Implement loading mechanisms

Examples:
- University of Nevada, Las Vegas: CONTENTdm → Islandora
- Arizona State University: Home-grown Django Repository → Islandora
- Fedora 3 → Fedora 4/5
“Major”: some aspect of the component is not backwards-compatible with the existing version

Reduced scale version of the complete system migration.

Examples:

- **Fedora 4/5→6**: Changed the storage layer from modeshape to the Oxford Common Filesystem Layout (OCFL) + SQL-based index
- **Hardware infrastructure**
  - **UNLV**: split-server→redundant single-server
  - **ASU**: Ansible + AWS Elastic Compute Cloud→Docker + AWS Elastic Container Services
Continuity of Service

When migrating between systems (or major system components) you can either:

- take the system/component offline during the update or
- switch to a redundant copy you created beforehand.

This question grows in significance with the size of your content corpus.
Object Oriented Migrations
Islandora @ Carnegie Mellon

Julia Corrin
Carnegie Mellon University

CNI 2022
Migration History

1994
HELIOS
1 collection
First of its kind to provide deep access to archival materials. Used NLP for search.

1999
DIVA
16 collections
The next generation of HELIOS, built and managed entirely by the CMU libraries.

2011
ArchivalWare
26 collections
A vended system, designed to let digitized content to be added without technical support.

2021
Islandora
Open source and heavily customized in house. Intended to regain control of content and features.
Primarily System/Feature Oriented

- Existing system feels “old” and “clunky”
- Improved interface design
- Feature enhancements
  - IIIF implementation
  - Mirador book viewer
  - Additional content types
Frequently Object Oriented

- Metadata
  - No standard metadata schema
  - No controlled vocabularies
  - Inconsistent field usage
  - Missing fields
  - Data formatting (eg. dates)

- File management
  - Missing master files
  - Duplicate and outdated files
  - Mismatched page and object counts
  - Potential reintroduction of redacted files

Metadata

Browse

Title

Collection
- H. John Heinz III

Series

Archival Topic
- Civil Rights

Folder Title
- Civil Rights Act of 1990 -- JH Working Files

Identifier
- \Heinz\box00326\fld00006\bdl0007\doc0002\Heinz_box00326\fld00006\bdl0007\doc0002.r

Rights

Type
- pdf

Thumbnail
- \Heinz\box00326\fld00006\bdl0007\doc0002\THUMBNAIL\Heinz_box00326\fld00006\bdl000

Document ID
- 734887
**Technical Debt**

<table>
<thead>
<tr>
<th>Collection Management</th>
<th>Decisions</th>
<th>Technical Debt</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing archival functions</strong> . . .</td>
<td>. . . require making decisions. Decision styles:</td>
<td>. . . result in varying different types and degrees of TD.</td>
<td>Debt accrual costs you in different ways, impacting execution of ongoing archival functions.</td>
</tr>
<tr>
<td>• Description</td>
<td><strong>Active/Deliberate</strong></td>
<td>• Non-standardized metadata</td>
<td>• Resource impact</td>
</tr>
<tr>
<td>• Access</td>
<td>• Strategic</td>
<td>• Poor UX</td>
<td>• Value impact</td>
</tr>
<tr>
<td>• Preserving context (relationships &amp; structure)</td>
<td>• Tactical</td>
<td>• Weak documentation</td>
<td>• Quality impact</td>
</tr>
<tr>
<td>• Preservation access</td>
<td><strong>Passive/Inadvertent</strong></td>
<td>• Work arounds vs. workflows</td>
<td></td>
</tr>
</tbody>
</table>
Object Based Technical Debt

First Migration
200,000 objects

Second Migration
300,000 objects

Third Migration
400,000 objects

- Non-standard metadata
- Incorrectly oriented pages
- Incorrectly mapped metadata fields

- Non-standard metadata
- Incorrectly oriented pages & duplicate scans
- Incorrectly mapped metadata fields

- Non-standard metadata, inconsistent between collections
- Incorrectly oriented pages & duplicate scans
- Missing master scans
- Incorrectly mapped metadata fields
- Missing metadata fields
### Objects vs. Items

<table>
<thead>
<tr>
<th>Objects</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>A complete document – eg. a book, newspaper, photograph, etc.</td>
<td>All the “things” that need to be migrated/assessed/reformatted</td>
</tr>
<tr>
<td>May include multiple data streams</td>
<td>Much, much more than the number of items in your repository</td>
</tr>
<tr>
<td>○ Metadata</td>
<td></td>
</tr>
<tr>
<td>○ Pages</td>
<td></td>
</tr>
<tr>
<td>○ Derivatives</td>
<td></td>
</tr>
<tr>
<td>○ Preservation Information</td>
<td></td>
</tr>
</tbody>
</table>

### Metadata Records
- ~400,000
- 2.75+ million

### Pages Total
- 1,040

### Shakespeare’s 3rd folio
- 7

### Average number of pages per document

<table>
<thead>
<tr>
<th>Metadata Records</th>
<th>Pages Total</th>
<th>Number of pages Shakespeare’s 3rd folio</th>
<th>Average number of pages per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>~400,000</td>
<td>2.75+ million</td>
<td>1,040</td>
<td>7</td>
</tr>
</tbody>
</table>
• Can’t rely on legacy system exports as a guide
  ○ Vended repository “ate” documents
  ○ Metadata exports did not include all items
  ○ Some items never made it from Diva (1999-2011) to ArchivalWare (2011-2021)
    ■ And when were items removed on purpose???
• Can’t use existing repository files as service copies
  ○ Previous repository relied on web optimized PDF-A for service copies
  ○ Quality of existing service copies is degraded due to compression
  ○ IIIF supports using TIFFs/JPEGs as service copies
• Can’t locate and/or can’t identify the master files
  ○ 25+ years of master files on tape back ups
  ○ No voting system across back ups
  ○ Original scans and rescans present for some documents
    ■ Eg. Scanned microfilm and scanned original for newspapers
• Can’t define completeness
  ○ Pages in PDFs don’t match the number of JPEGs found

Unanticipated Challenges
Object Oriented Technical Debt Remediation

- Masters and derivatives are now both managed via the repository
  - Still working to eliminate rescanned content
- Internally consistent metadata schema
  - Some custom metadata fields were still required
  - EDTF date implementation
- Authority file implementation and URI inclusion
- Reversion to TIFF and JPEG masters
  - PDF-A copies still available as a derivative, but not longer used as service master
TL;DR:

Significant object based technical debt directly affected our ability to achieve the goals of our migration:

A feature rich repository
Does this help migrate Fedora 2?

Yes

Do It!

No

Drop It!

END
How It Started

- UMD’s Digital Collections launched in Fedora 2 in 2005
- Digital programs expanded to include large-scale audiovisual digitization projects in the following decade
- Digital A/V content was stored in a vendor-based streaming media service (Sharestream) and accessed via Fedora 2 metadata records

How It (Was) Going

- Sharestream/Fedora 2 process was inefficient and user-hostile
- UMD Libraries brought up our Fedora 4–based repository in 2016, and we re-engaged with the Fedora community
- Fedora 2 badly needed to be sunset as our primary repository
- Began a year-long Avalon pilot, implemented in 2019
Starting with Values

**Open**
Academy-owned, open-source infrastructure is core to our approach

**Sustainable**
Our business is permanence, and need systems that will grow with our program

**Usable**
Our research methods incorporated interviews, site visits, and accessibility review

**Inclusive**
We employed co-creation techniques to engage commonly excluded stakeholders in selection
Landing on Avalon

User stories
Generated from interviews and observations

Requirements
Met by out of the box functionality

Essential Issues
Required custom development
This is a quarter of our project plan, which definitely went as expected.
- **Strategy**: Bring up minimum viable instance to meet deadlines in a large grant funded project, use to stress test the application prior to full migration
- Prepared initial ingest of **1,199** videos from the *Liz Lerman Dance Exchange* project.
- Launched Avalon in production mode on May 4, 2021
...And Finding Out

- At our media repository scale, we could not use Avalon to store and deliver preservation files as we had initially hoped.
- Asset transcoding at scale would require weeks of buffer time for collection ingests.
- Group access control management and roles for Avalon would not work as planned with our Grouper configuration.
- Would need to build much more sophisticated file download and request fulfillment features to work with Aeon and various departments.
- Target collection mapping proved to be one of the most time-consuming initial activities.
- No single “source of truth” for location of assets and relevant access control rules.
10,600

A/V files to migrate in 6 months.
Given those challenges, argued for “Cleared Decks” levels of focus for the central migration team for ~6 straight months.
With the Product

Built an external IP Manager service and token–URL based Request Fulfillment feature

With the Content

Re–generated, manually downloaded, and pulled access files from hard drives (but avoided the binder of CDs!)

With the Metadata

Cross–walked custom descriptive metadata schema to Avalon’s ingest format; fully re–mapped source collections
Technical Labor

- Variable custom development required
- Binary and metadata management always presents new surprises
- New workflows need to be developed, stress-tested, and documented by stakeholders
- Grappling with decades of technical debt and evolving standards (ask me about legacy filenames!) is an unavoidable headache

Emotional Labor

- Software, systems, and workflows have emotional effects on participants
- Change leadership is challenging
- Communication plans must be empathetic but keep participants well informed
- Team leads have to listen, hype, coach, troubleshoot, and occasionally debate
- High turnover rates affect the team
And then we migrated the rest of our digital collections out of Fedora 2 without any problems at all.

The end.
Our Message:

- Our shared stories can provide experience and expertise to help guide migration decisions
- Don’t wait until it’s too late – make migration planning part roadmap planning
- Data migrations affect everyone
- Collaboration and communication with all stakeholders is key
Questions?

Seth Shaw – seth.e.shaw@asu.edu
Julia Corrin – jcorrin@andrew.cmu.edu
Kate Dohe – katedohe@umd.edu
Arran Griffith – arran.griffith@lyrasis.org