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

Arizona State University

Complete & Partial System Migrations

Seth Shaw, Digital Library Software Developer

Migrations

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

Migrations

Complete System Migrations: Moving house

System Component “Major” Upgrades: Remodeling the kitchen

Complete System Migration

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

System Component “Major” Upgrades

“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

Migrations

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 before-hand.

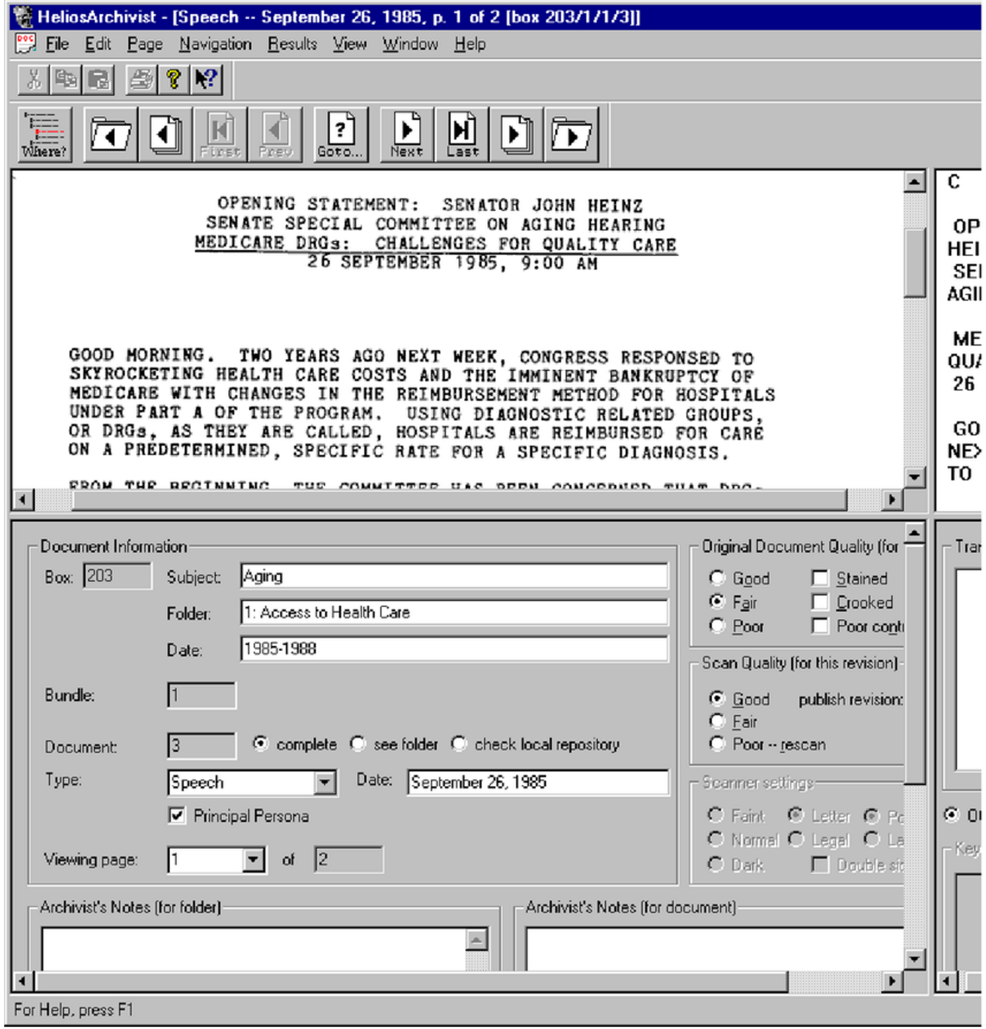
This question grows in significance with the size of your content corpus.

Object Oriented Migrations

Islandora @ Carnegie Mellon

Julia Corrin
Carnegie Mellon University

CNI 2022



OPENING STATEMENT: SENATOR JOHN HEINZ
SENATE SPECIAL COMMITTEE ON AGING HEARING
MEDICARE DRGs: CHALLENGES FOR QUALITY CARE
26 SEPTEMBER 1985, 9:00 AM

GOOD MORNING. TWO YEARS AGO NEXT WEEK, CONGRESS RESPONSED TO SKYROCKETING HEALTH CARE COSTS AND THE IMMINENT BANKRUPTCY OF MEDICARE WITH CHANGES IN THE REIMBURSEMENT METHOD FOR HOSPITALS UNDER PART A OF THE PROGRAM. USING DIAGNOSTIC RELATED GROUPS, OR DRGs, AS THEY ARE CALLED, HOSPITALS ARE REIMBURSED FOR CARE ON A PREDETERMINED, SPECIFIC RATE FOR A SPECIFIC DIAGNOSIS.

FROM THE BEGINNING, THE COMMITTEE HAS BEEN CONCERNED THAT DRG-

Document Information

Box: 203
Subject: Aging
Folder: 1: Access to Health Care
Date: 1985-1988
Bundle: 1
Document: 3 ● complete ○ see folder ○ check local repository
Type: Speech Date: September 26, 1985
☑ Principal Persona
Viewing page: 1 of 2

Original Document Quality (for

○ Good □ Stained
● Fair □ Crooked
○ Poor □ Poor cont

Scan Quality (for this revision)

● Good publish revision:
○ Fair
○ Poor -- rescan

Scanner settings

○ Faint ● Letter ● Po
○ Normal ○ Legal ○ La
○ Dark □ Double si

Archivist's Notes (for folder)

Archivist's Notes (for document)

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

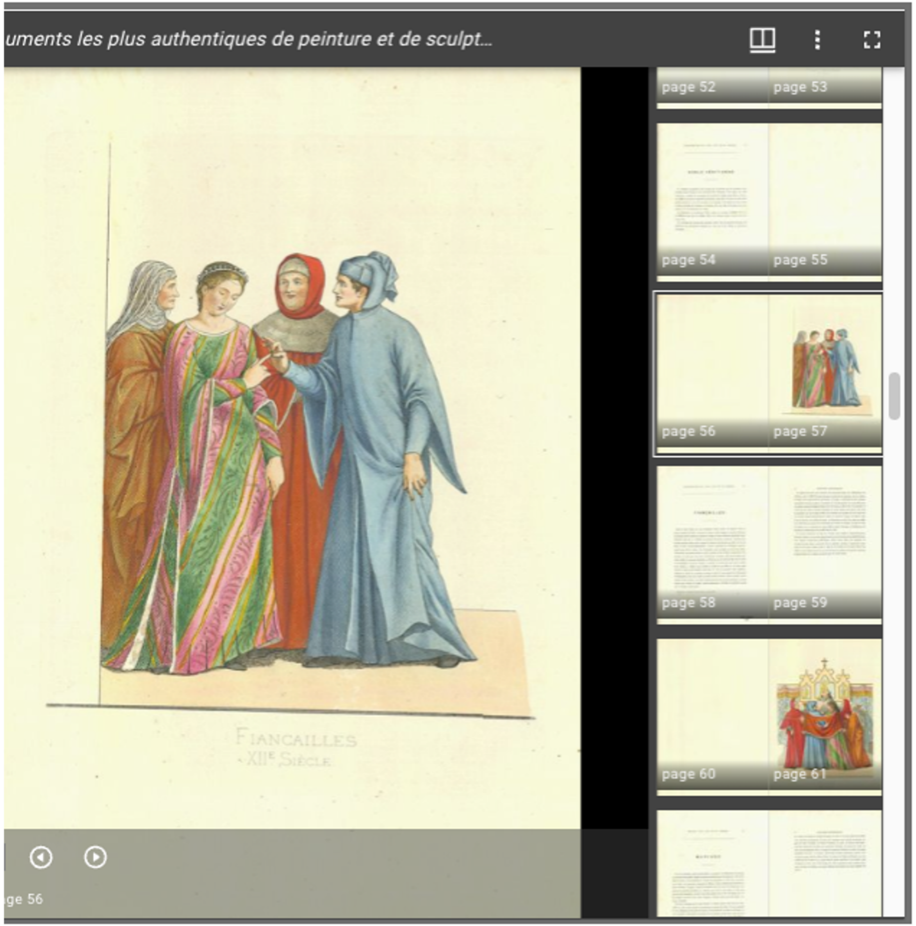
Islandora

Open source and heavily customized in house. Intended to regain control of content and features.

Migration Goals

Primarily System/Feature Oriented

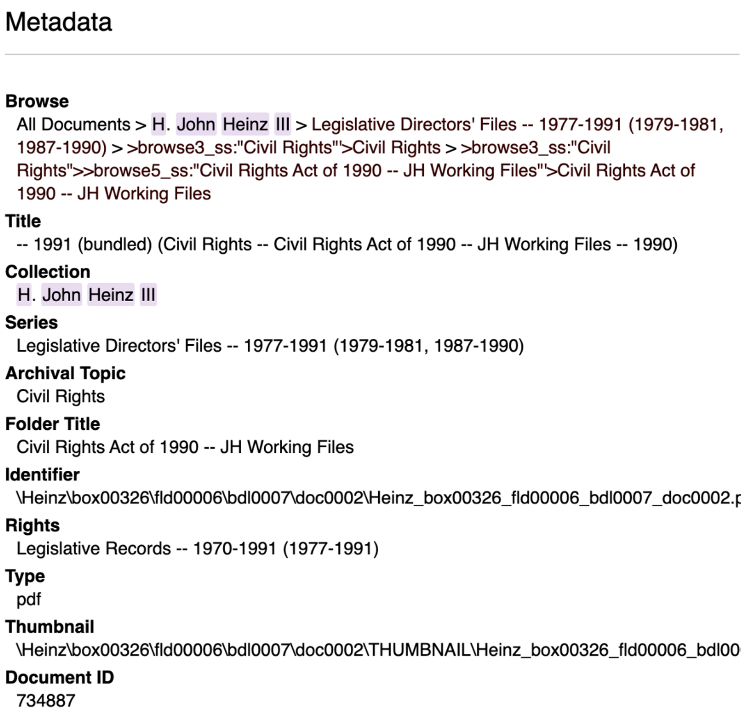
- Existing system feels “old” and “clunky”
- Improved interface design
- Feature enhancements
  - IIIF implementation
  - Mirador book viewer
  - Additional content types



Migration Challenges

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
All Documents > H. John Heinz III > Legislative Directors' Files -- 1977-1991 (1979-1981, 1987-1990) > >browse3_ss:"Civil Rights">Civil Rights > >browse3_ss:"Civil Rights">>browse5_ss:"Civil Rights Act of 1990 -- JH Working Files">Civil Rights Act of 1990 -- JH Working Files

Title
-- 1991 (bundled) (Civil Rights -- Civil Rights Act of 1990 -- JH Working Files -- 1990)

Collection
H. John Heinz III

Series
Legislative Directors' Files -- 1977-1991 (1979-1981, 1987-1990)

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.p

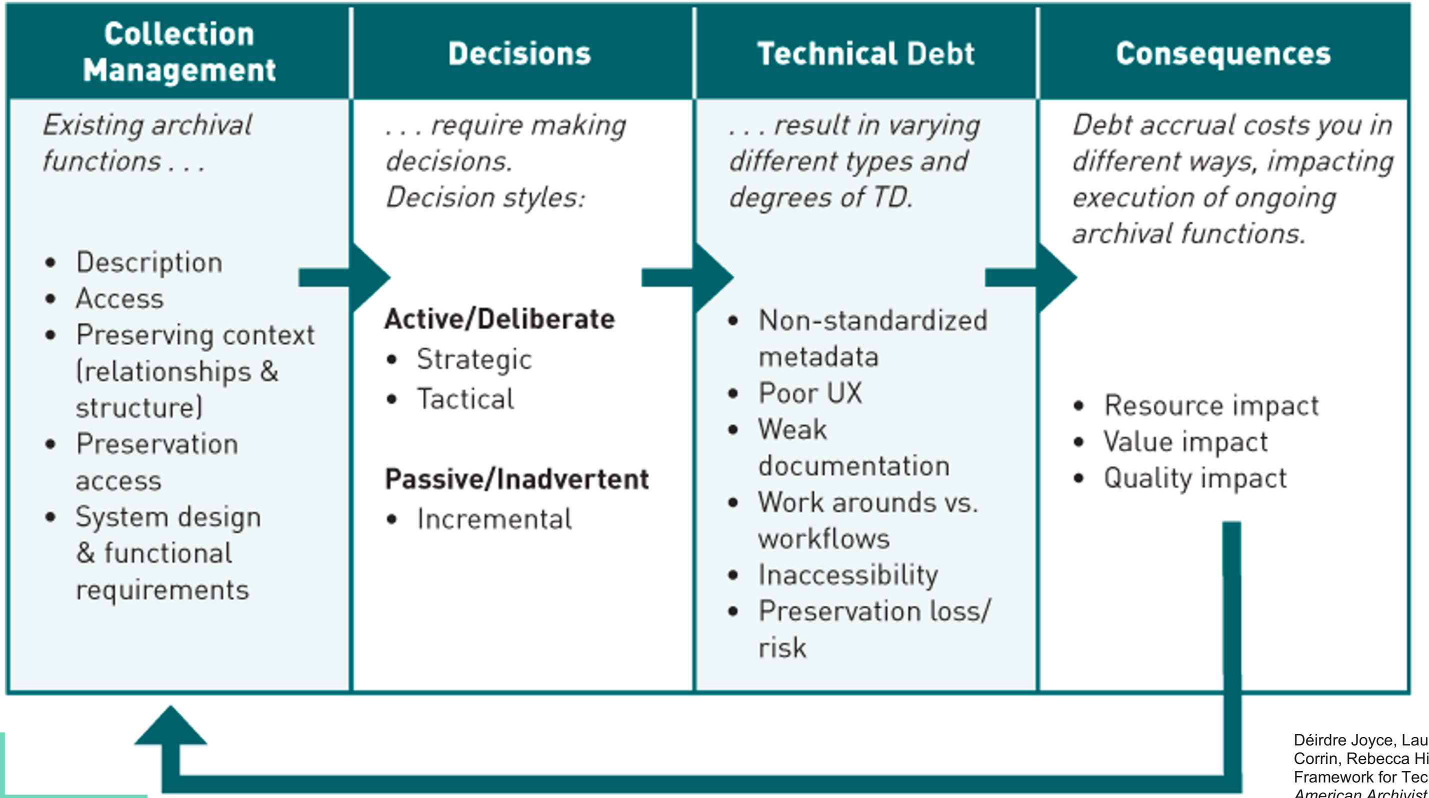
Rights
Legislative Records -- 1970-1991 (1977-1991)

Type
pdf

Thumbnail
\Heinz\box00326\fld00006\bdl0007\doc0002\THUMBNAIL\Heinz_box00326_fld00006_bdl00

Document ID
734887

Technical Debt



Collection Management

Existing archival functions . . .

- Description
- Access
- Preserving context (relationships & structure)
- Preservation access
- System design & functional requirements

Decisions

. . . require making decisions.
Decision styles:

Active/Deliberate

- Strategic
- Tactical

Passive/Inadvertent

- Incremental

Technical Debt

. . . result in varying different types and degrees of TD.

- Non-standardized metadata
- Poor UX
- Weak documentation
- Work arounds vs. workflows
- Inaccessibility
- Preservation loss/ risk

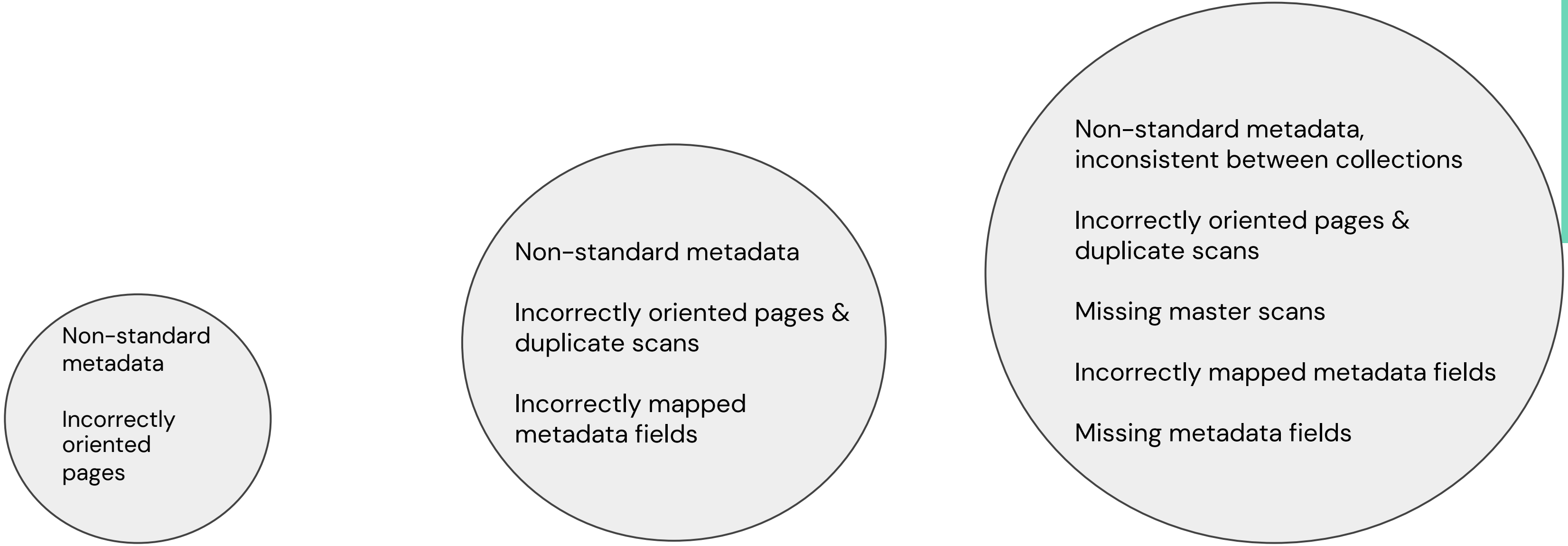
Consequences

Debt accrual costs you in different ways, impacting execution of ongoing archival functions.

- Resource impact
- Value impact
- Quality impact

Déirdre Joyce, Laurel McPhee, Rita Johnston, Julia Corrin, Rebecca Hirsch; Toward a Conceptual Framework for Technical Debt in Archives. *The American Archivist* 1 March 2022; 85 (1): 104–125. doi: https://doi.org/10.17723/2327-9702-85.1.104

Object Based Technical Debt



Non-standard metadata

Incorrectly oriented pages

First Migration
200,000 objects

Non-standard metadata

Incorrectly oriented pages & duplicate scans

Incorrectly mapped metadata fields

Second Migration
300,000 objects

Non-standard metadata, inconsistent between collections

Incorrectly oriented pages & duplicate scans

Missing master scans

Incorrectly mapped metadata fields

Missing metadata fields

Third Migration
400,000 objects

Objects vs. Items

~400,000	2.75+ million	1,040	7
Metadata Records	Pages Total	Number of pages Shakespeare’s 3rd folio	Average number of pages per document

Objects

- A complete document - eg. a book, newspaper, photograph, etc.
- May include multiple data streams
  - Metadata
  - Pages
  - Derivatives
  - Preservation Information

Items

- All the “things” that need to be migrated/assessed/reformatted
- Much, much more than the number of items in your repository

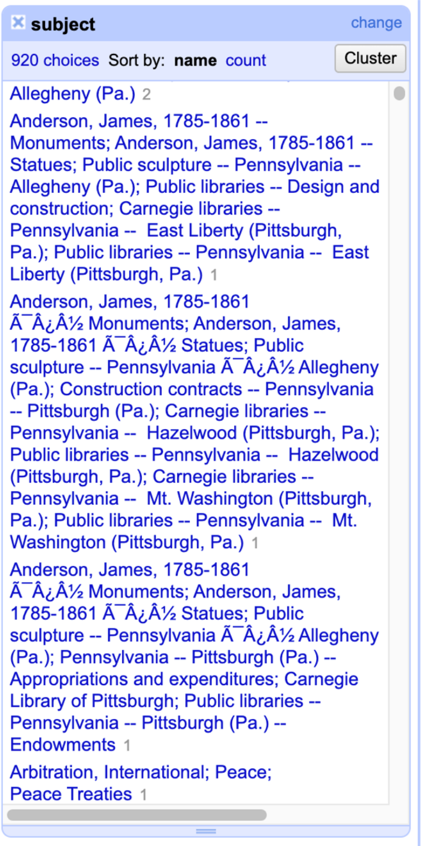
Unanticipated Challenges

- **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

Migration Outcomes

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



Allegheny (Pa.) 2

Anderson, James, 1785-1861 -- Monuments; Anderson, James, 1785-1861 -- Statues; Public sculpture -- Pennsylvania -- Allegheny (Pa.); Public libraries -- Design and construction; Carnegie libraries -- Pennsylvania -- East Liberty (Pittsburgh, Pa.); Public libraries -- Pennsylvania -- East Liberty (Pittsburgh, Pa.) 1

Anderson, James, 1785-1861 Ã¯Â¿Â½ Monuments; Anderson, James, 1785-1861 Ã¯Â¿Â½ Statues; Public sculpture -- Pennsylvania Ã¯Â¿Â½ Allegheny (Pa.); Construction contracts -- Pennsylvania -- Pittsburgh (Pa.); Carnegie libraries -- Pennsylvania -- Hazelwood (Pittsburgh, Pa.); Public libraries -- Pennsylvania -- Hazelwood (Pittsburgh, Pa.); Carnegie libraries -- Pennsylvania -- Mt. Washington (Pittsburgh, Pa.); Public libraries -- Pennsylvania -- Mt. Washington (Pittsburgh, Pa.) 1

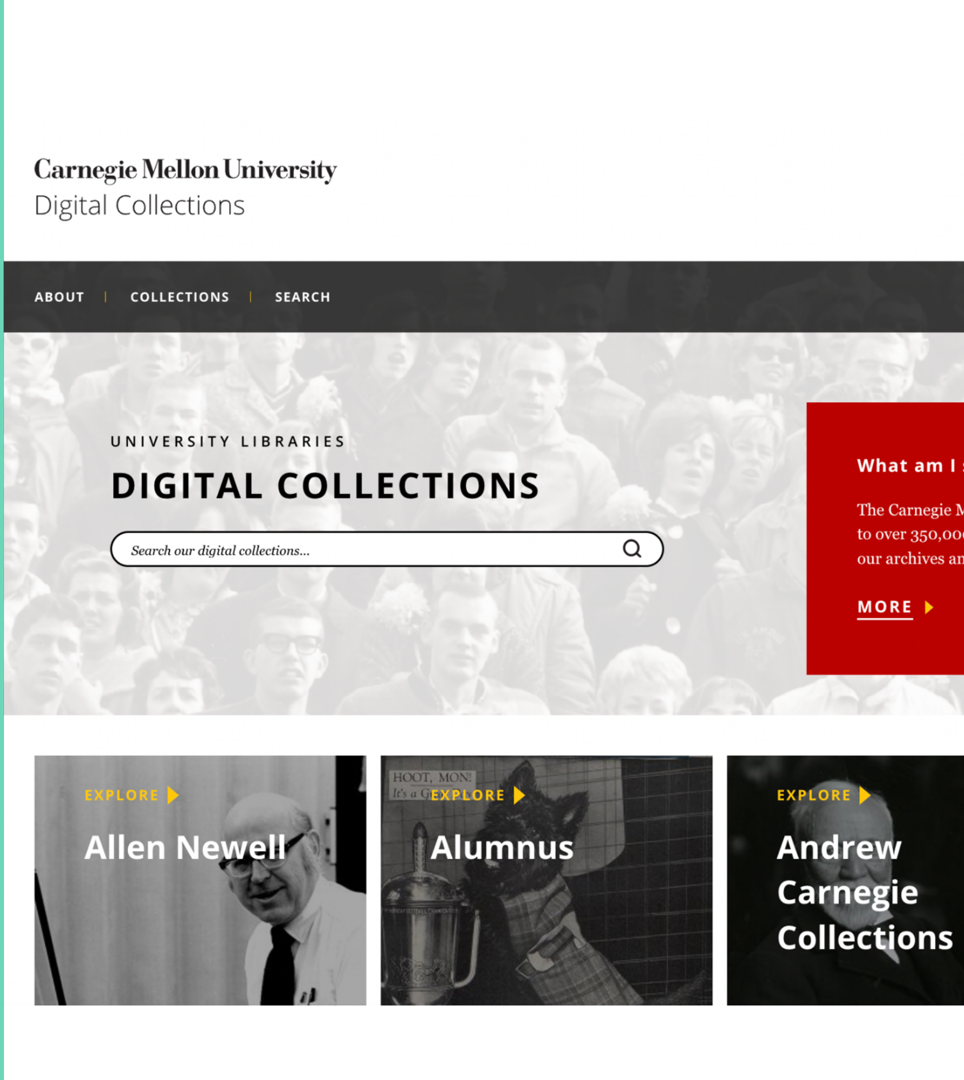
Anderson, James, 1785-1861 Ã¯Â¿Â½ Monuments; Anderson, James, 1785-1861 Ã¯Â¿Â½ Statues; Public sculpture -- Pennsylvania Ã¯Â¿Â½ Allegheny (Pa.); Pennsylvania -- Pittsburgh (Pa.) -- Appropriations and expenditures; Carnegie Library of Pittsburgh; Public libraries -- Pennsylvania -- Pittsburgh (Pa.) -- Endowments 1

Arbitration, International; Peace; Peace Treaties 1

TL;DR:

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

A feature rich repository



Carnegie Mellon University
Digital Collections

ABOUT | COLLECTIONS | SEARCH

UNIVERSITY LIBRARIES

DIGITAL COLLECTIONS

Search our digital collections...

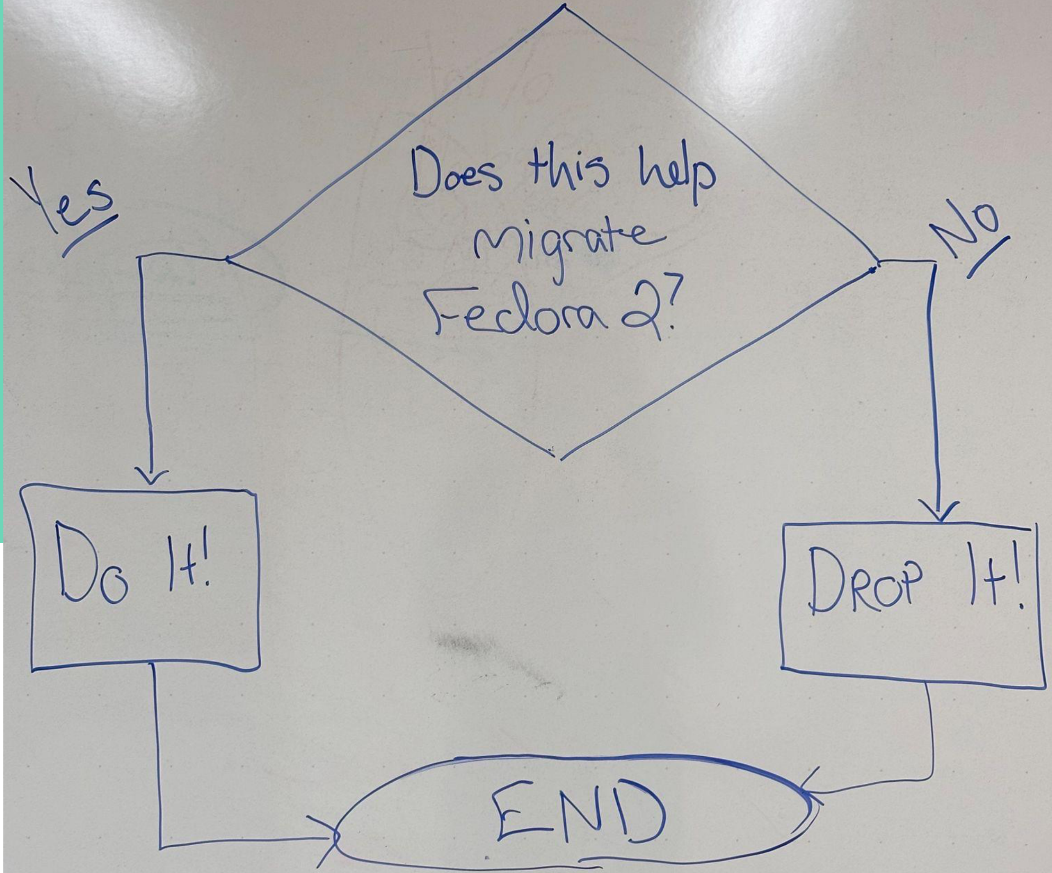
What am I

The Carnegie M
to over 350,00
our archives an

MORE
EXPLORE
Allen Newell
HOOT, MON!
It's a G
EXPLORE
Alumnus
EXPLORE
Andrew Carnegie Collections

Avalon
@ UMD Libraries

Kate Dohe
Director of Digital Programs & Initiatives



Ground Truths

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

Fifty

User stories

Generated from interviews and observations

Twenty Five

Requirements

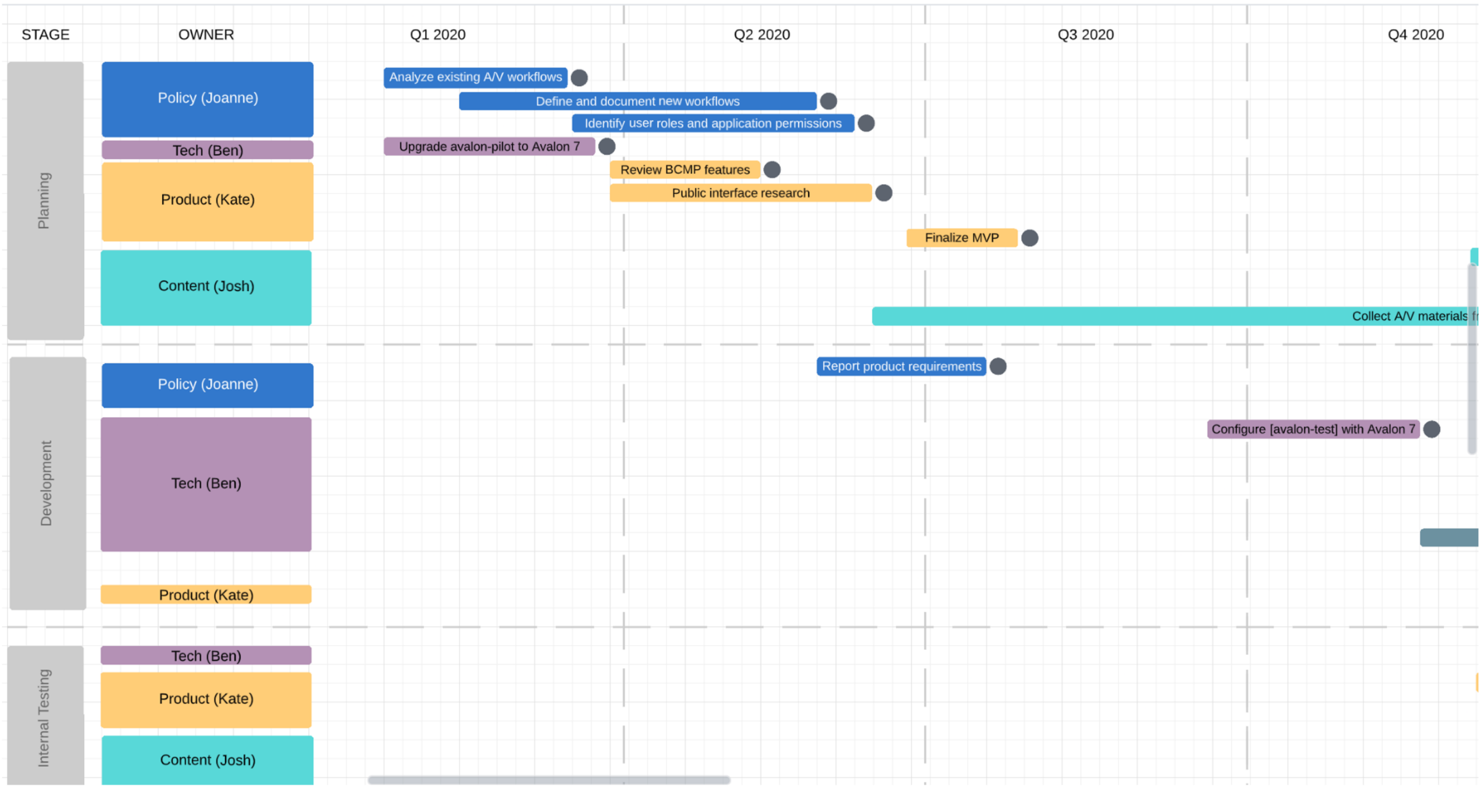
Met by out of the box functionality

Eight

Essential Issues

Required custom development

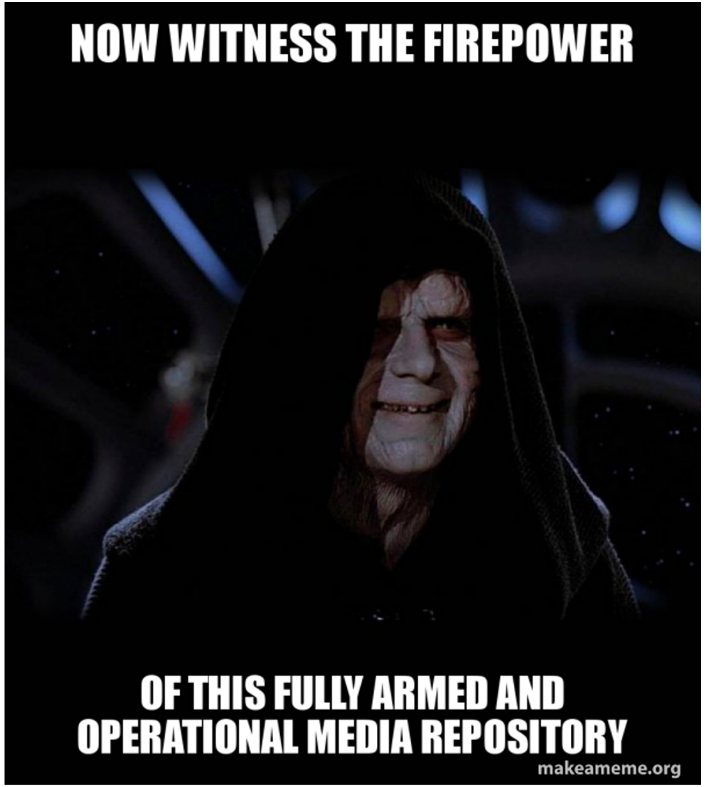
Here comes the hard part



This is a quarter of our project plan, which definitely went as expected.

MVP Launch

- **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 Exchance* project.
- Launched Avalon in production mode on May 4, 2021


NOW WITNESS THE FIREPOWER
OF THIS FULLY ARMED AND OPERATIONAL MEDIA REPOSITORY
makeameme.org

Outcomes from Launch

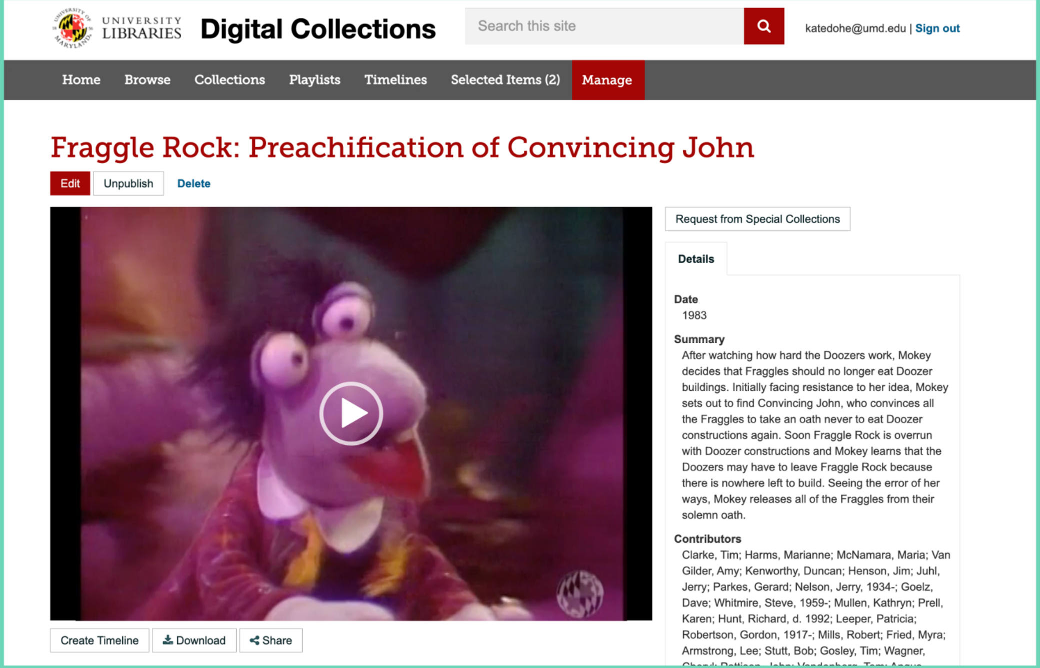
…And Finding Out

- 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.

Listen to Convincing John



Fraggle Rock: Preachification of Convincing John

Date
1983

Summary
After watching how hard the Doozers work, Mokey decides that Fraggles should no longer eat Doozer buildings. Initially facing resistance to her idea, Mokey sets out to find Convincing John, who convinces all the Fraggles to take an oath never to eat Doozer constructions again. Soon Fraggle Rock is overrun with Doozer constructions and Mokey learns that the Doozers may have to leave Fraggle Rock because there is nowhere left to build. Seeing the error of her ways, Mokey releases all of the Fraggles from their solemn oath.

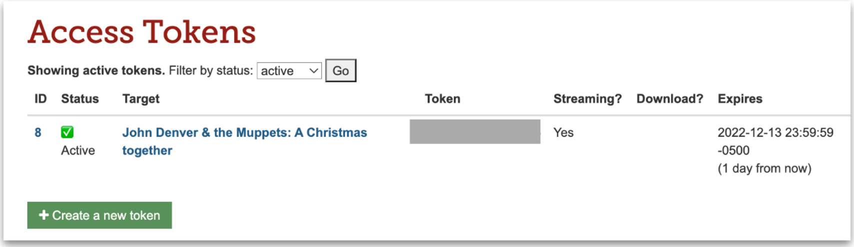
Contributors
Clarke, Tim; Harms, Marianne; McNamara, Maria; Van Gilder, Amy; Kenworthy, Duncan; Henson, Jim; Juhl, Jerry; Parkes, Gerard; Nelson, Jerry, 1934-; Goelz, Dave; Whitmire, Steve, 1959-; Mullen, Kathryn; Prell, Karen; Hunt, Richard, d. 1992; Leeper, Patricia; Robertson, Gordon, 1917-; Mills, Robert; Fried, Myra; Armstrong, Lee; Stutt, Bob; Gosley, Tim; Wagner, Cheryl; Pattison, John; Vandenberg, Tom; Angus

Given those challenges, argued for “Cleared Decks” levels of focus for the central migration team for ~6 straight months.

Getting There…

With the Product

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



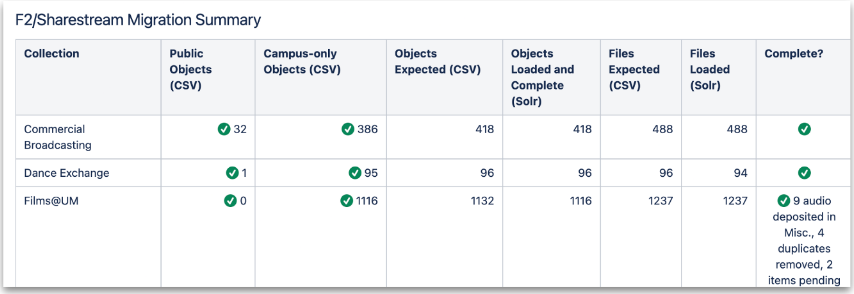
Access Tokens

Showing active tokens. Filter by status: active Go

ID	Status	Target	Token	Streaming?	Download?	Expires
8	✓ Active	John Denver & the Muppets: A Christmas together	[redacted]	Yes		2022-12-13 23:59:59 -0500 (1 day from now)

With the Content

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

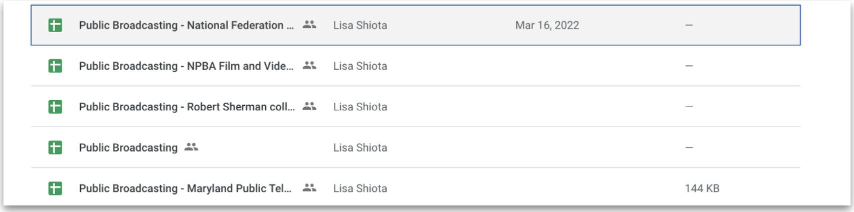


F2/Sharestream Migration Summary

Collection	Public Objects (CSV)	Campus-only Objects (CSV)	Objects Expected (CSV)	Objects Loaded and Complete (Solr)	Files Expected (CSV)	Files Loaded (Solr)	Complete?
Commercial Broadcasting	✓ 32	✓ 386	418	418	488	488	✓
Dance Exchange	✓ 1	✓ 95	96	96	96	94	✓
Films@UM	✓ 0	✓ 1116	1132	1116	1237	1237	✓ 9 audio deposited in Misc., 4 duplicates removed, 2 items pending

With the Metadata

Cross-walked custom descriptive metadata schema to Avalon's ingest format; fully re-mapped source collections



Public Broadcasting - National Federation ...	Lisa Shiota	Mar 16, 2022	–
Public Broadcasting - NPBA Film and Vide...	Lisa Shiota		–
Public Broadcasting - Robert Sherman coll...	Lisa Shiota		–
Public Broadcasting	Lisa Shiota		–
Public Broadcasting - Maryland Public Tel...	Lisa Shiota		144 KB

Migrations are…

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.

Summary

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

Carnegie Mellon University

University of Maryland 18 56

ASU Arizona State University

Fedora™