Approaching Big Challenges with Small Steps

Audio/Video Digitization and Preservation at NYPL

Setup

- Intros (Me, NYPL, What we do)
- Digital preservation often starts with digitization
- 2005-2015 - Stage One: Ready or not, it’s happening
- 2015-2017 - Stage Two: Big ambitions and some momentum
- 2017-2018 - Stage Three: Reality checks
- 2018- - Stage Four: Born-digital challenges

About Me

Nick Krabbenhoeft (crabbin’ hofft is OK)
Digital Preservation Manager, New York Public Library
Visiting Assistant Professor, Pratt Institute (NYC)

nickkrabbenhoeft@nypl.org
Twitter, Slack, Github, etc: @nkrabben

NYPL

- 88 branches in the Bronx, Manhattan, and Staten Island
- 4 research centers
  - Stephen A Schwarzman Building
  - New York Public Library for the Performing Arts
  - Schomburg Center for Research in Black Culture
  - Science, Industry, and Business Library

NYPL Digital Special Collections

- 14 million digitized images (2.4 TIFF preservation files)
- 150,000 born-digital archival files described in finding aids (250+ file formats)
- 50-100 digital recordings of dance, theater, and other performances per year
- 4 PB of digitized audio and moving image (AMI) media

NYPL AMI Digitization Initiative

- In the 2020s, magnetic media will become increasingly inaccessible
  - Deteriorating media
  - Increasingly rare playback equipment
  - Retiring playback engineers
- Digitization is the best action to continue preserving these materials
  - The additional cost digitization and digital preservation is lower than the cost of inaction, the complete loss of collection material. https://coi.weareavp.com
- Across 3 research libraries, NYPL has ~250,000 rare or unique recordings on VHS, open-reel, CD, U-matic, DAT, film, and other media types

Digital preservation often starts with digitization

- Digitization is not digital preservation.
- After digitization you still have to answer questions like:
  - Where are the files?
  - Are they the files I think they are?
  - Have they changed?
  - Can I access them?
- We can make digital preservation easier by setting the right conditions

Digital Collection Lifetime Costs

Based on a David Rosenthal (LOCKSS) rule of thumb

- 50% of costs go to ingest
  - Digitizing or transferring files
  - Quality Control
  - Creating descriptions
  - Matching files to descriptions
- 33% of costs go to storage
  - Buying and replacing storage hardware
  - Checking fixity
- 16% of costs go to access
  - Building access systems
  - Moving materials from storage to access

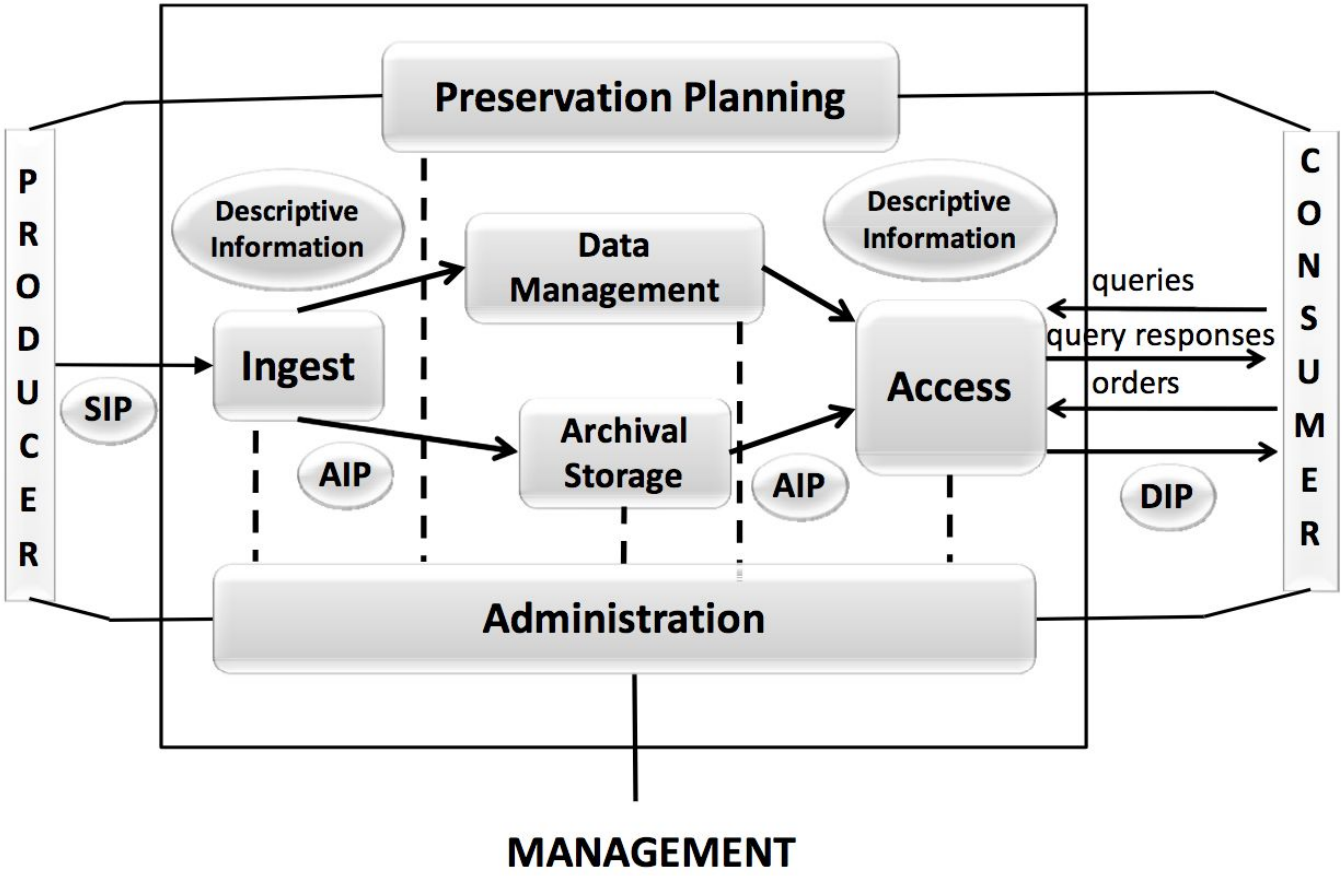


Figure 4-1: OAIS Functional Entities

Reduce Ingest Costs to Reduce Preservation Costs

- Information Package
  - File specifications
    - File format, codec, signal quality
  - Metadata specifications
    - Field names, Required Fields, Controlled vocabulary
  - Fixity specifications
    - Checksum type, Checksum storage
  - Folder specifications
    - File names, Folder names, Folder hierarchy
- Processes
  - Quality assurance
    - Are packages meeting the specifications we set?
  - AIP Generation
    - What extra work do we have to do to be ready for preservation?

File Specifications: Wrapper/Codec

Audio and Video files are defined by their wrapper and their codec

- Wrapper: the structure that contains the streams of media
- Codec: the specific way a stream of media is encoded

Wrapper/Codec

- Wrapper: 35 mm film
- Stream 1 (Light blue)
  - SDDS Digital audio
- Stream 2 (between left sprockets)
  - Dolby Digital audio
- Stream 3 (left of picture)
  - Stereo analog audio
- Stream 4 (picture)
  - Anamorphic video

Wrapper/Codec

Common Digitization Wrappers/Codecs

- WAV/PCM
- MP3/MP3

- MOV/V210
- MKV/FFV1
- MXF/JPEG2000
- MP4/H264

Metadata: Possible Carriers

- Spreadsheets
  - Easy to read
  - Easy to edit by hand
  - Hard to validate

Asset							Process
File		Bibliographic Item				Notes	Recording
Reference Filename	File Role	Acquisition ID*	Division code*	Class mark/ID*	Title*	Access Note*	Location
myd_mgzidvd57822	Preservation Master		myd	*MGZIDVD 5-7822	ODFY_Tere		
myd_mgzidvd57822	Preservation Master		myd	*MGZIDVD 5-7822	ODFY_Tere		
myd_mgzidvd57822	Preservation Master		myd	*MGZIDVD 5-7822	ODFY_Tere		
myd_mgzidvd57822	Edit Master		myd	*MGZIDVD 5-7822	ODFY_Tere		
myd_mgzidvd57822_v026250244_sc.mp4	Edit Master		myd	*MGZIDVD 5-7822	ODFY_Tere O'Connor		
myd_mgzidvd57822_v026250245_sc.mp4	Edit Master		myd	*MGZIDVD 5-7822	ODFY_Tere O'Connor		

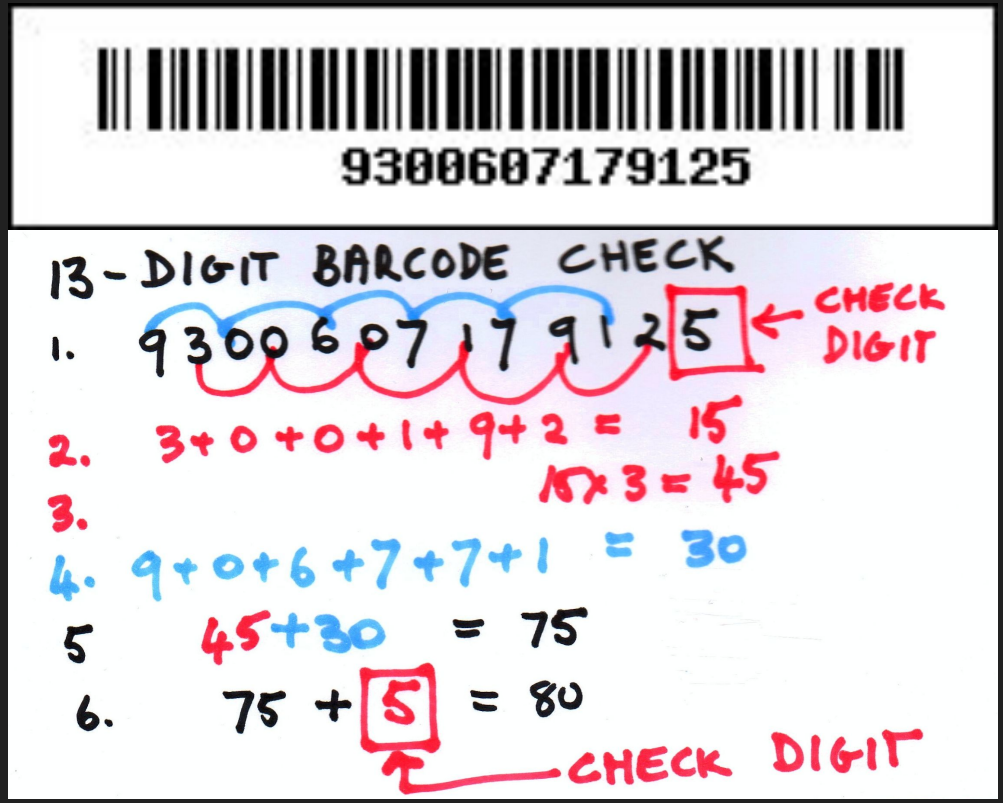
Metadata: Possible Carriers

- Spreadsheets
  - Easier to read
  - Easier to edit by hand
  - Harder to validate
- Text data formats (XML, JSON, etc)
  - Harder to read
  - Easier to edit with code
  - Easier to validate (with schema)


```
"asset": {
    "referenceFilename": "myh_abc123_v
    "fileRole": "pm",
    "schemaVersion": "2.0.0"
},
"bibliographic": {
    "primaryID": "abc123",
    "classmark": "*ABC 123",
    "formerClassmark": "*NCOW 123",
    "nonCMSID": "string123",
    "cmsItemID": "123456",
    "cmsCollectionID": "string123",
    "catalogBNumber": "B12345678",
    "mssID": "12345",
    "barcode": "33433123456789",
    "divisionCode": "myd",
    "vernacularDivisionCode": "DAN",
    "projectCode": "projectXYZ",
    "title": "Dance Audio 1",
    "date": "Late 1970s",
```

Fixity

- Information used to make sure your data hasn't changed
- Barcodes often include a "check digit" to ensure the other numbers are correct



Fixity: Choices

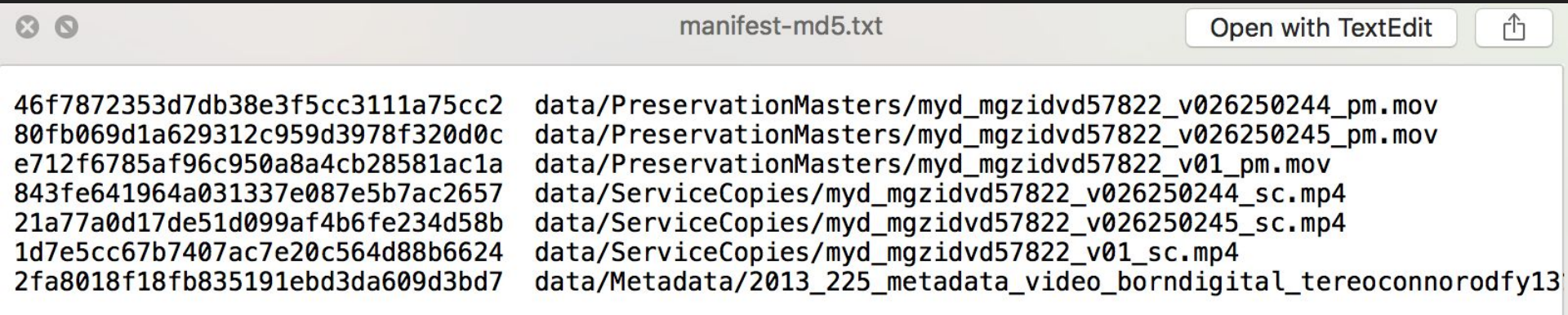
- We use a checksum based on an algorithm (MD5, SHA-1, SHA-2,…)

Fixity: Choices

- We use a checksum based on an algorithm (MD5, SHA-1, SHA-2,…)
- Fixity has to be stored somewhere
  - In its own file (filename.md5)

Fixity: Choices

- We use a checksum based on an algorithm (MD5, SHA-1, SHA-2,…)
- Fixity has to be stored somewhere
  - In its own file (filename.md5)
  - In a manifest (folder_md5.txt)



Fixity: Choices

- We use a checksum based on an algorithm (MD5, SHA-1, SHA-2,…)
- Fixity has to be stored somewhere
  - In its own file (filename.md5)
  - In a manifest (folder_md5.txt)
  - In metadata (<checksum>…</checksum>)

```
"technical": {
    "checksum": "337c9226768608945a5804904a389d04"
    "filename": "abc123",
    "extension": "wav",
```

Reduce Ingest Costs to Reduce Preservation Costs

- Information Package
  - File specifications
    - File format, codec, signal quality
  - Metadata specifications
    - Field names, Required Fields, Controlled vocabulary
  - Fixity specifications
    - Checksum type, Checksum storage
  - Folder specifications
    - File names, Folder names, Folder hierarchy
- Processes
  - Quality assurance
    - Are packages meeting the specifications we set?
  - AIP Generation
    - What extra work do we have to do to be ready for preservation?

Stage 1: Ready or not, it’s happening

2005-2015

- Projects decided collection-by-collection
- Desire, but no mandate for preservation

Information Package

- Uncompressed MOV/V210 and WAV/PCM
- Spreadsheets from collections, XML exports from engineers, and MediaInfo exports
- MD5 sidecars or manifests
- Received hard drive with 1 project folder with 1 folder for preservation masters and 1 folder for service copies

Package Walkthrough

Quality Control

- File specs
- Metadata specs
- Fixity specs
- Folder specs

Quality Control

- File specs (By hand)
- Metadata specs
- Fixity specs
- Folder specs

Quality Control

- File specs (By hand)
- Metadata specs (By hand)
- Fixity specs
- Folder specs

Quality Control

- File specs (By hand)
- Metadata specs (By hand)
- Fixity specs (By hand)
- Folder specs

Quality Control

- File specs (By hand)
- Metadata specs (By hand)
- Fixity specs (By hand)
- Folder specs (By hand)

Critique the Package

Generate the AIP

- Create technical metadata
- Combine all metadata into spreadsheet
- Move/rename folders
- Move checksums to bag manifest

Workflow Challenges during Stage 1

- A lot of time spent manipulating metadata
- Very hard to implement quality control work
- Bags were getting larger as projects got larger (80+ TB)

Stage 2: Big ambitions and some momentum

2015-2017

- Funding to digitize 100,000+ objects
- Mandate for preservation

Information Package

- Uncompressed MOV/v210 and WAV/PCM
- Structured metadata
- BagIt fixity information
- 1 bag per digitized item

Package Walkthrough

Quality Control

- File specs - Commercial System (To be purchased)
- Metadata specs (Custom metadata schema)
- Fixity specs (BagIt Packaging)
- Folder specs (To be determined)

Metadata Walkthrough

- What is a schema?
  - A document defining all the rules for your metadata

Metadata Walkthrough

- What is a schema?
- How did NYPL build a schema?
  - Built a custom schema based on our own metadata requirements

Metadata Walkthrough

- What is a schema?
- How did NYPL build a schema?
- What's a better option for a schema?
  - Use a shared schema like PBCore

Generate the AIP

- Fix bad metadata
- Delete unwanted files
- Fix bad packaging (folder names)

Workflow Challenges during Stage 2

- Too many required fields for metadata
- Our custom metadata scheme was probably too custom
- Quality control software was hung up in purchasing
- Not enough storage space for all of our files
- If fixity information didn’t match, we didn’t know where the damage was

Stage 3: Reality checks

2017-2018

- Don’t have all expected resources
- Do have feedback from vendors

Information Package

- Uncompressed WAV/PCM and losslessly compressed MKV/FFV1
- Structured metadata with relaxed requirements
- BagIt fixity information and FFV1 fixity information
- 1 bag per digitized item

Package Walkthrough

Quality Control

- File specs - Mediaconch and QCTools (Open Source)
- Metadata specs - Structured metadata schema
- Fixity specs - BagIt Packaging
- Folder specs - Custom script

Lossless Compression

Using a codec that represents the same amount of information with less bytes. However, you need to have a standard way to decompress the bytes.

Lossless Compression

Using a codec that represents the same amount of information with less bytes. However, you need to have a standard way to decompress the bytes.

- NYPL is a losslessly compressed version of New York Public Library

Lossless Compression

Using a codec that represents the same amount of information with less bytes. However, you need to have a standard way to decompress the bytes.

- NYPL is a losslessly compressed version of New York Public Library
- SASB is a losslessly compressed version of Stephen A. Schwartzmann Building, where I work

Common Lossless Compression Methods

- MOV/V210
  - Uncompressed 10-bit color transfer NTSC SD video = 100 GB/hr
- MXF/JPEG2000
  - Same quality = 40-50 GB/hr
  - Users include Library of Congress
- MKV/FFV1
  - Same quality = 40-50 GB/hr
  - Users include City University of New York, Austrian Media Archives, Irish Film Archives, NYPL

The Case for Lossless Compression

- If you lose a single byte of compressed data, you lose more information.
  - For example, if NYPL looked like NXPL
- Because of that libraries like uncompressed data.
  1 bit of damage = 1 bit of information loss
- But you generally, don’t lose bits, you lose files or hard drives or servers
- If you’re lossless files are 50% the size, you can afford to keep 2x as many copies of your lossless files.

MediaConch Demo

Measures files against a policy of specifications

MediaConch Demo

1. Copy a policy

MediaConch Demo

1. Copy a policy
2. Edit a rule

MediaConch Demo

1. Copy a policy
2. Edit a rule
3. Test files

Generate the AIP

- Fix bad metadata
  - Less because of reduced metadata requirements
- Delete unwanted files
  - Less because of write blockers

Workflow Challenges during Stage 3

- New media types challenge our assumptions
  - Film
  - Born-Digital Video
- Small vendors can’t meet all of our requirements

Stage 4: Born-digital challenges

2017-

- What do we do with born-digital video?
- How do we manage output from 20-40 different videographers?

Information Package

- Lossy compressed MOV/ProRes Files?
- Back to spreadsheets?
- BagIt fixity information
- 1 bag per recorded performance

Package Walkthrough

Quality Control

- File specs - Mediaconch and QCTools (Open Source)
- Metadata specs - TBD
- Fixity specs - BagIt Packaging
- Folder specs - Custom script

Generate the AIP

- Fix bad metadata
  - More because using spreadsheets again
- Delete unwanted files
  - More because of many smaller vendors
- Fix bad packaging
  - More because of many smaller vendors

If we could start fresh...

Information Package

- MKV/FFV1 and WAV/PCM
- PBCore XML metadata
- BagIt fixity information
- 1 bag per item

A lot of vendors can produce this.

Small Steps

1. You don’t have to fix everything at once
2. Document your packages and workflows
3. Work on the biggest headaches first
4. Use what others have put out their
5. Make more changes when you have the opportunity

Resources

Cost of Inaction - https://coi.weareavp.com/
PBCore - http://pbcore.org/
BagIt - https://patchbay.tech/2017/12/15/using-bagit-in-2018/
MediaConch - https://mediaarea.net/MediaConch/Documentation/HowToUse

General Resources
AMIA Open Workflows - https://github.com/amiaopensource/open-workflows
Guidelines on Audio Preservation - IASA TC-04
Guidelins on Video Preservation - IASA TC-06