Management of Digital Programs in Special Collections

Robin C. Flick
Manages Digital Conversion and Media Reformattting
University of Maryland College Park Libraries
CUA CLSC747, July 19, 2017
With my background in music and music technology, I focused on sound and moving image archives and digital preservation during graduate school.
Overview

- Digitization Production
  - In-house, outsourced
- Project Planning
- Standards
- Ingesting/Archiving
  - Access, copyright
- Digital Preservation

Robin C. Fike, CLSC747, Monday, July 18, 2016
Digital Collections at UMD Libraries

- Office of Digital Collections and Research under the Dean of Libraries
  - 2005-2006: Fedora Commons selected at architecture
  - First “Best Practices” released in 2007
  - Working on a revision

- Digital Collections within Special Collections
  - 2010, decentralization of staff and resources

- 2012, re-centralization under Digital Systems and Stewardship (IT)

- Digital Conversion and Media Reformating
  - Expansion into additional formats, services, outsourcing

- Digital Programs and Initiatives
  - Expansion into additional digital programs, repository development, digital preservation

- Difference between UMD model and other libraries

The University of Maryland Libraries made a decision in 2005 to create a digital repository using the open source Fedora digital repository. At the time, other solutions, such as CONTENTdm, were not as fully realized, and the University of Maryland Libraries wanted a scalable and flexible solution that could enable the long-term preservation of digital objects. Granted the programming and librarian support necessary to develop such a system - at the time, the staffing of the then Office of Digital Collections and Research (DCR) included an Assistant Dean, two digital librarians, one database administrator, one developer, one user interface specialist, and the time of a metadata librarian - DCR embarked on a long-term development project, designed to be completed in stages and to meet a number of needs throughout the Libraries.

Early DCR made move to create Best Practices. Haven’t been updated in three years, need to be re-examined. The Best Practices document, however, has been a useful crutch for us, especially with turnover in the department.

It provides guidance and documentation relevant to understanding the motivation behind the creation of digital collections. It includes a broad definition of “digital master” which is something we still haven’t quite worked out. (“best copy available). The Best Practices touch on copyright considerations, metadata, selection, project management and staffing.

Importance of documentation, authenticity. Best practices can change!!!

2010: one manager and four students. Developers moved to ITD
About DCMR

- **Mission:**
  - The Digital Conversion and Media Reformating Department (DCMR) seeks to support the Libraries’ collection development goals and strategic priorities for preservation and access by working with collection managers and subject specialists to digitize collections of all formats through a centralized, production-based environment.

- **Role:**
  - Provide the digitization operations for the seven College Park Libraries, and serve as a leader for digitization in the community. Outside of Special Collections department.

- **Well-established program**
  - Like many of the digitization programs, now at point where refining and revising current guidelines and practices, especially what can be done in-house
  - Standardized workflows across collection areas, projects
  - Upscaled digitization capacity by outsourcing

- **Current Capacity**
  - In-house digitization: text, image, audio requests and small projects
  - Outsourced: large projects, audio and moving image requests

Robin C. Pike, CLSC747, Monday, July 18, 2016
I manage general operations which includes project planning, funding, budgeting, all the bureaucratic paperwork, creating policies and procedures, and acting as the vendor liaison for projects across the seven campus libraries. I also do outreach to the campus librarians and staff, and through the subject liaisons, campus outreach. In addition to managing large outsourced projects and general operations, I manage two full-time librarians, one of which is dedicated to managing the NEH National Digital Newspaper Program grant for the state of MD, the other of which manages our in-house digitization center in Hornbake Library, and the requests and projects done in the center. I also manage a half-time graduate assistant, who has been working on our processes for handling and preserving born-digital collections (one of the articles you read). We have approximately 11 hourly students to support in-house digitization for requests and small projects, material prep for outsourced projects, quality assurance processes for in-house and vendor projects, outreach efforts (Pinterest and Twitter), and general support.
Examined:

- What can you feasibly and most cost-effectively do?
- What is better to do in-house?
- What is better to do through a vendor?
- What is better to out-source?

Out-source: Special Collections using a graphic designer to create graphics for their biggest exhibits
Hornbake Digitization Center (HDC)

- In-house operations
  - 4 flatbed stations
  - 1 planetary station
  - 2 digital audio workstations
  - 1 digital video workstation (in development)
  - 2 staff workstations
Talk about plan for setting up new digitization center

HDC Equipment/Software

- Digital Imaging
  - Epson Expression 10000XL
  - Epson Perfection V700
  - Zeutschel planetary scanner
  - Canon digital camera
  - SilverFast
  - Adobe Photoshop

- Digital Audio
  - Variety of legacy players
  - A-D converter, patchbay
  - Adobe Audition
Historic Maryland Newspapers Project

- 4 staff workstations
- Project Manager
- 2 student collation prep and QR
- 1 microfilm inspection
Digital Collections

- Fedora-based Repository: digital.lib.umd.edu
  - 30,840 metadata records (many with multiple images per record)
  - 16,479 still image records
  - 1,078 EAD finding aids
  - 8,127 books
  - 5,288 audio, video, film records
  - 81 TEI-Encoded documents
- Internet Archive: 7,295 books
- HathiTrust: 600 books
- Chronicling America: est. 211,263 pages
- Est. +150TB

Robin C. Pike, CLIS 747, Monday, July 18, 2016
Through the previous six years, both pre-DIC and during, digitization has generally trended upward with some fluctuation, particularly with special projects, such as the American Archive project in FY13, a project for which we received external funding, and the digitizing the School of Music recordings in FY16.
Lifecycle of Digitization in Special Collections

- Selection
- Description
- Planning
- Digitization
- Quality Assurance
- Ingest
- Archiving
- Preservation
Selection

- Primarily done by the librarians or curators since they know their content best
- Projects and requests
- Money
  - Grant
  - Donor
  - Trust/Private foundation
- Copyright, or lack of
- Preservation
- Access
Planning Metadata

- What do you want to describe?
- What is your audience?
- How will your users be using the collections?
- What are you planned access points?
Description—Schemas

- MODS/METS
  - http://www.loc.gov/standards/mods/
  - http://www.loc.gov/standards/mets/

- Dublin Core
  - http://dublincore.org/

- PREMIS
  - http://www.loc.gov/standards/premis/
Other Metadata Standards

- VRA Core (describing works of visual arts)
- ATMC (Audio Technical Metadata)
  - From IU/Harvard Sound Directions
- AudioMD and VideoMD (for use with MODS/METS)
  - http://www.loc.gov/standards/amdvmmd/
- PBCore (Public Broadcasting Metadata Dictionary, for use with Dublin Core)
- reVTMD (Video Technical Metadata from NARA)
  - http://www.archives.gov/preservation/products/reVTMD.xsd

Because imaging metadata standards usually aren’t enough for audio and moving image media
Embedded Metadata Standards

- FADGI:
  - Basic Guidelines for Minimal Descriptive Metadata in Digital Still Images
  - Embedding Metadata in Digital Audio Files
- Library of Congress worked on BWFMetaEdit
- NARA worked on AVIMetaEdit, and video embedded standards are upcoming

*Description*: Identifier [comma space] type [comma space] comment [semicolon-space if more than one identifier]

*If no labeling*: Identifier

*Originator*: Country code and OCLC University Code, separated by a comma and a space. “US, UMC”

*Originator Reference*:

CC Country code: (2 characters) based on the ISO 3166-1 standard [2]

OOO Organisation code: (3 characters) based on the EBU facility codes, Tech 3279 [3].

NNNNNNNNNNNNN Serial number: (12 characters extracted from the recorder model and serial number) This should identify the machine’s type and serial number.

HHMMSS OriginationTime (6 characters,) from the <OriginationTime> field of the BWF.

These elements should be sufficient to identify a particular recording in a human-useful form in conjunction with other sources of information, formal and informal. In addition, the USID contains:

RRRRRRRRRR Random Number (9 characters 0-9) Generated locally by the recorder using some reasonably random algorithm.

*Origination Date*: Date of digitization. ISO standard. Should be auto-generated.


*Time Reference (translated)*: Time code. Should be auto-generated.
Description—Going Further

- Repurposing
  - Using data in finding aids
  - Cross-walking MARC
  - Importing spreadsheets as CSV files
  - Exporting records in databases into CSV
- Harvesting (OAI-PMH)
- Linked data
  - http://linkeddata.org/
- Metatags
  - http://www.w3schools.com/tags/tag_meta.asp
Metadata Practice

- UMDM, UMAM, METS
  - Founded early when few schemas existed, do not use standardized schema
- Expanding into embedded metadata for audio and video
- What are our plans for the future?
  - Fedora 4
    - Linked data (schema.org)
    - MODS/METS
    - PBCore
    - MARCXML

Robin C. Pike, CLSC747, Monday, July 18, 2016
Digitization Initiatives
Committee/Project Planning

- Projects approved by Associate Deans
- Project proposal prioritization (proposal form and rubric)
  - Project evaluation and ranking
    - Example
- Finds/allocates funding
- Budget approved by Libraries administration
- Start on formal project plans
  - Example

Robin C. Pike, CLSC747, Monday, July 18, 2016
Other Funding

- Grants
  - NEH
    - National Digital Newspaper Program
    - Humanities Collections and Reference Resources
  - Mellon Grant


- [https://docs.google.com/spreadsheets/d/1hFaexQevrca72XPytUPELudu3R0145r8DMGuUGRugm4/edit?usp=sharing](https://docs.google.com/spreadsheets/d/1hFaexQevrca72XPytUPELudu3R0145r8DMGuUGRugm4/edit?usp=sharing)
<table>
<thead>
<tr>
<th>FY18 Projects</th>
<th>Division</th>
<th>Amount</th>
<th>Accounts/Partners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hebraica</td>
<td>PSD/R&amp;L</td>
<td>$17,700</td>
<td>Gift/Jewish Studies Dept.</td>
</tr>
<tr>
<td>Godowsky Collection</td>
<td>PSD/IPAM</td>
<td>$3,400</td>
<td>Gift</td>
</tr>
<tr>
<td>Woody’s Children</td>
<td>PSD/SCPA</td>
<td>$10,000</td>
<td>Gift</td>
</tr>
<tr>
<td>American Bandmasters</td>
<td>PSD/SCPA</td>
<td>$800</td>
<td>Am. Bandmasters Assoc.</td>
</tr>
<tr>
<td>Association</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarice Smith Digital Video</td>
<td>PSD/SCPA</td>
<td>$18,000</td>
<td>Gift/The Clarice</td>
</tr>
<tr>
<td>Vinegar Syndrome Films</td>
<td>PSD/LMS</td>
<td>$13,000</td>
<td>Operating</td>
</tr>
<tr>
<td>Prange Posters</td>
<td>CSS/Prange</td>
<td>$10,000</td>
<td>Gifts (3)</td>
</tr>
<tr>
<td>MMC Serials</td>
<td>CSS/SCUA</td>
<td>$10,000</td>
<td>Gift/Library of Am. Broad.</td>
</tr>
<tr>
<td>MPT</td>
<td>CSS/SCUA</td>
<td>$8,000</td>
<td>Operating/MPT</td>
</tr>
<tr>
<td>Diamondback Newspaper</td>
<td>CSS/SCUA</td>
<td>&lt;$56,000</td>
<td>UMD Launch Campaign</td>
</tr>
<tr>
<td>Student Newspapers</td>
<td>CSS/SCUA</td>
<td>$6,000</td>
<td>Operating</td>
</tr>
<tr>
<td>Katherine Anne Porter</td>
<td>CSS/SCUA</td>
<td>$3,000</td>
<td>KAP Literary Trust</td>
</tr>
<tr>
<td>Spiro Agnew</td>
<td>CSS/SCUA</td>
<td>$9,100</td>
<td>Gift</td>
</tr>
<tr>
<td>Athletics Video</td>
<td>CSS/SCUA</td>
<td>$4,000</td>
<td>Gift</td>
</tr>
<tr>
<td>Gymkana Photographs</td>
<td>CSS/SCUA</td>
<td>$10-15K</td>
<td>UMD Launch/SPH</td>
</tr>
</tbody>
</table>
## Project Planning Gantt Chart

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task 1</td>
<td>Task 2</td>
<td>Task 3</td>
<td>Task 4</td>
<td>Task 5</td>
<td>Task 6</td>
<td>Task 7</td>
<td>Task 8</td>
<td>Task 9</td>
<td>Task 10</td>
<td>Task 11</td>
<td>Task 12</td>
<td>Task 13</td>
<td>Task 14</td>
<td>Task 15</td>
<td>Task 16</td>
<td>Task 17</td>
</tr>
</tbody>
</table>

[Project Planning Gantt Chart Image]

Robin C. Piko, CLSC747, Monday, July 18, 2016
Working with Digitization Vendors

- Define the scope of the project
- Define the standards you want to use
  - Digitization
  - Metadata
  - File format and quality
- Define timeline of project
- Define stakeholders of the project and what you need from them and when
Choosing a Vendor

- Ask regional institutions for regional suggestions
- Ask for recommendations on professional listservs
  - ALA, SLA, SAA, ARSC, AMIA, etc.
  - Some associations have regional lists posted
- Ask similar institutions
- Network at conferences
- If you haven’t worked with the vendor before, ask for references for similar projects

Robin C. Pike, CLSC747, Monday, July 18, 2016
Planning Projects Through Vendors

- Get vendor to deliver a statement of work (SOW) from them describing what they do and how they’ll do it—negotiation between you and vendor
- SOW should define the deliverables and how you will receive them
- If project is more than a few items, get the vendor to agree to deliver a pilot project, approve this
- Start the full project and plan for a delivery schedule
- Example: Technical Specifications (from a SOW)

Robin C. Piko, CLSIC747, Monday, July 18, 2016
Digitization Standards

- Digitization
  - File formats [Example: Technical Specifications]
  - FADGI: best practice but might not be practical
- Guidelines:
- Resources:
  - [http://www.digitizationguidelines.gov/resources/](http://www.digitizationguidelines.gov/resources/)
  - Still Image Working Group
  - Audio Visual Working Group

Robin C. Pike, CLSC747, Monday, July 18, 2016
Other Projects

- CDL [California Digital Library]
  - http://www.cdlib.org/services/collections/massdig/
  - Many Library Partners have developed their own digitization standards

- NCEcho
  - Digitization and Metadata Guidelines

- IU Bloomington Media Preservation Initiative
  - Meeting the Challenge of Media Preservation:
What is the Right Format?

- What are you trying to capture?
- How will the files be used?
- What is your storage capacity?
- Can your institution support non-carrier digital formats?
- What software/hardware do you have?
- What software/hardware might you support in the future?

Generally:
- Widely-adopted proprietary format or opensource format, cross-platform

What software/hardware might you support in the future?
You don’t want to choose software that produces a specific file format that might not be around in the future—additional migration=extra work.
Generally: for example AVI was a windows format that can now be read on a Mac. It has been around for a while, and has been adopted by many institutions as an acceptable archival digital video format. The H.264 codec for MPEG video files is a widely accepted codec, which is used on Blue-ray discs, as well as popular sites and applications like YouTube, Vimeo, iTunes, and Adobe Flashplayer MXF/J2K at LC
Like I mentioned, scanned photos comprise the second set of digital photos. They should be scanned on a high-quality scanner at at least 300dpi when blown up to an 8x10, though the actual resolution scanned at may vary. As the diagram explains, 300 dpi is good enough for 8x10 prints, but if you’re scanning 4x6 prints or slides or negatives, you’ll want a much higher resolution if you want to print the image.

300dpi is the archival standard, but lately, especially with smaller prints, I have been scanning at 600 dpi. This produces a much larger file, but it is at a much higher quality (much less pixilated), which is beneficial if you have the room to save these large files.

We also scan in 24-bit color, not grayscale, which includes the context of the paper and the variance of ink in the image.
Current Text Standards

- Book or Serial Publication
  - Archival: Uncompressed TIFF or lossless compressed JPEG2000 at color depth and pixilation appropriate for application (3000 pixels along long edge, 8”x10” 300 dpi equivalent)
  - Derivative formats dependent on use, most likely JPEG or PDF
  - HathiTrust has separate set of guidelines

- Newspapers
  - Grayscale raster formats for masters, PDFs and OCR text for access and use
  - NDNP has separate set of guidelines for images and metadata
  - TIE (Text Encoding Initiative): SGML, XML

Robin C. Pike, CLSC747, Monday, July 18, 2016
The difference in the master formats accounts for sound recordings that are lower quality or more complex. High-quality recordings that contain voice only are sufficiently digitized at 48 kHz.
The difference in the master formats accounts for sound recordings that are lower quality or more complex. High-quality recordings that contain voice only are sufficiently digitized at 48 kHz.

<table>
<thead>
<tr>
<th>File</th>
<th>Container</th>
<th>Extension</th>
<th>Resolution</th>
<th>Bitdepth</th>
<th>Chroma Subsampling</th>
<th>Frame Rate</th>
<th>Timecode</th>
<th>Audio Channels</th>
<th>Audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Master</td>
<td>Uncompressed Quicktime File Format</td>
<td>.mov</td>
<td>Native</td>
<td>10bit</td>
<td>4:2:2</td>
<td>Native</td>
<td>Original</td>
<td>PCM, 48kHz, 16bit</td>
<td></td>
</tr>
<tr>
<td>Streaming</td>
<td>MPEG-4, H.264</td>
<td>.mp4</td>
<td>Native</td>
<td>8-bit</td>
<td>4:2:2</td>
<td>Native</td>
<td>Original</td>
<td>AAC, 48 kHz, 256kbps</td>
<td></td>
</tr>
</tbody>
</table>

Also receive metadata "sidecar" file
Why?

- Uncompressed standard for imaging
  - Archival
  - Proven
  - Lossless suits needs for Internet Archive

- Uncompressed standard for audio
  - Archival
  - Proven
  - Higher standard for music or mixed than spoken word

- Various standards for moving image
  - Various levels of projects and required deliverables
  - Growing support

Films at UM vs archival film footage
Quality Assurance

- Checking to make sure what you wanted to receive from the vendors or produced in-house is actually what you have
  - File naming
  - File formats
  - Number of files
  - Quality technical specifications
  - Checking deterioration v. artifacts created during digitization process (prominent in a/v)
    - AV Artifact Atlas:
      http://avga.bavc.org/artifactatlas/index.php/A/V_Artifact_Atlas

Robin C. Pike, CLSC747, Monday, July 18, 2016
Ingest/Archiving

- In-house
  - Usually one digital object created at one time
  - ILL, SCUA workflow for public domain books to IA: quarterly batch ingest
  - Monthly archiving

- Out-sourced
  - Generally per project ingests.
  - Fedora4 migration halting some ingests from 2015-2017
    - Will be ingesting in batch
    - Per project or monthly archiving, depending
Digital Preservation

- NDIIPP (National Digital Information Infrastructure and Preservation Program)
  - Sustainability of Digital Formats Planning for Library of Congress Collections
    - The Signal Blog
    - Various reports and sponsored studies

- TDR: not achievable in-house, but use established

- Bit storage: file storage

- Enhanced capabilities
  - Fixity checks, migration, normalization, etc.

- Current:
  - Multiple copies on spinning disk and tape.
  - Exploring memberships in APtrust, DPN, DuraSpace, and other possibilities of cloud/tape storage

Robin C. Falk, CLSC747, Monday, July 18, 2016
Campus restrictions and dark archive is most common with audiovisual materials, and we are currently working with our General Counsel lawyers to modify the policy and take a more progressive stance on access, making take-down policies standard across collections.
Copyright

- Public domain laws—does not apply to a/v
  - Audiovisual needs to be public within known rights, licensing

- http://www.librarycopyright.net/resources/digitalslider/

- Orphan works: http://copyright.gov/orphan/

- Mass-digitization:
  http://copyright.gov/docs/massdigitization/

- Library exceptions (Section 108):
  http://www.copyright.gov/title17/92chap1.html#108

- Fair Use (Section 107):
  http://www.copyright.gov/fairll102.html
Digital Collections Available:

- http://digital.lib.umd.edu/
- http://chroniclingamerica.loc.gov/newspapers/?state=Maryland&pid=umd:233169
Questions?

Robin C. Pike
Manager, Digital Conversion and Media Reformatting
UMD Libraries
rpike@umd.edu
Blog: https://dssumd.wordpress.com/
Twitter: @UMDDigitization
Facebook, Twitter, Instagram: @HistoricMDNews