Format Migrations at Harvard Library
General Framework and Plan Development

Project Description

This project through the National Digital Stewardship Resident program involved designing a format migration framework for obsolete digital formats in the digital repository. The format migration framework strategizes how to manage and execute migration projects and documents the general process for preparing for and performing a format migration, including but not limited to the steps that need to be taken, the decisions that need to be made, key stakeholders to include, the types of research and testing that needs to be done, migration artifacts that should be preserved, and templates to facilitate this process. The framework was developed by working through several real use cases with the Library's Preservation Services staff: Kodak PhotoCD (the primary case study for this poster), RealAudio, and SMIL audio playlists.

Overall Migration Framework

The most desired outcome from this project was a framework that could be used to develop a migration plan for any format in the repository. These are the essential components of the framework.

Project Workflow Example: Kodak Photo CD

The framework is used to guide the creation of a format-specific workflow

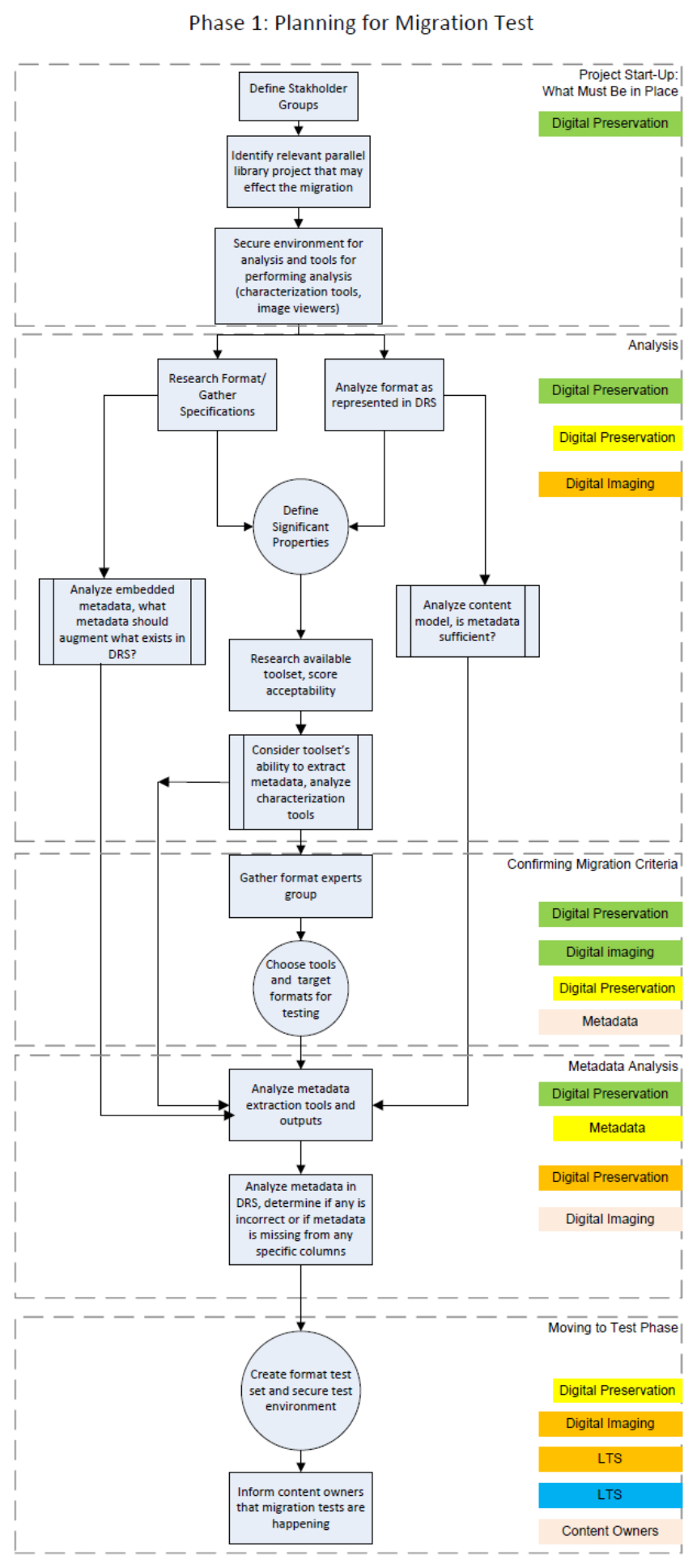
Deliverables and Metrics: Kodak Photo CD

Key deliverables are identified at each step of the framework to serve as evidence of the decisions made along the process and to gain consensus across stakeholder groups.

Phase 1: Planning for Migration Test

Component	Activity
Project Start-Up: What Must Be in Place	Define stakeholder groups and inform of project initialization. Refer to framework and any relevant past projects to serve as guidance for setting up the project. Refer to general data management/migration plans for a sense of how to engage constituents and when (DataCave, NSW).
	Identify other parallel library projects that might impact the timing or functionality of the migration plan (e.g. changes to database, batch ingest processing tools, metadata schemas, etc.)
	Create a protected environment for necessary testing tools, ensure that all platform requirements are satisfied in order to adequately analyze the content (e.g. text editors, image viewers, etc.) Additionally, have a secure way to pull content from the repository and for accessing unique applications or software (obsolete or in some cases Harvard-specific) that reduces risks of virus or malware.

Component	Activity
Analysis	Research format specifications and additional literature around recommended tools, target format, and processes for successful migration. Distill essential significant properties.
	Analyze content in the DRS grouped by shared technical characteristics. Distill into a plan for performing tests based on definitive groupings of content within a format (useful metadata to consider might be Roles, Methodology, Relationships, and specific technical metadata). Additional questions to potentially address in analysis: • How were the materials originally digitized • What was the chain of custody (donor deposits, Harvard owners, vendors, ingest of digital material, etc. • Any specific purpose or intent behind the objects to determine the appropriate preservation actions
Confirming Migration Criteria	Gather format expert groups and proposal for defining significant properties and migration criteria, confirm goals for successfully migrating content
Metadata Analysis	Assess existing metadata for content and identify any modifications that need to occur. Select appropriate characterization tools and compare against FITS output. Begin to assess any augmentation that needs to occur with the current schema and any migration-related metadata that should be added, for example a PREMIS event.
Moving into Test Phase	Determine the necessary OS, hardware, applications and other system requirements for the task. Secure the test environment and develop a strategy for creating a test set of content based on grouping, a variety of custom settings to apply for comparison, and metrics for performing quality assurance.

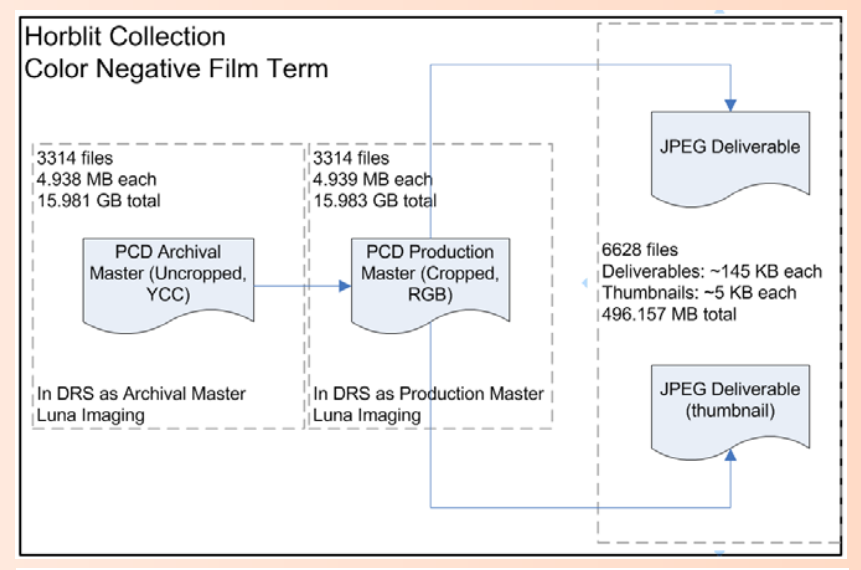


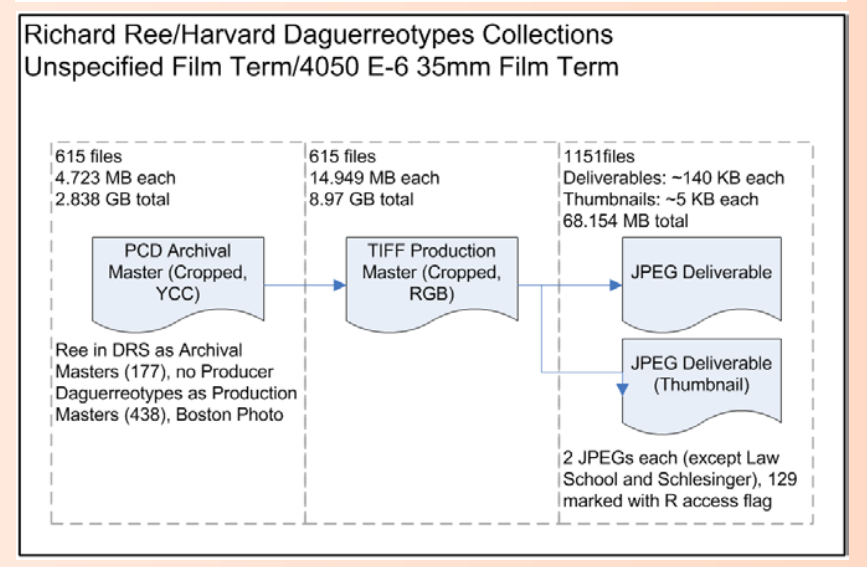
Stakeholder Involvement – Define Essential Roles Across the Project

- **Project Management** (Digital Preservation)
- **Technical Guidance/Format Expertise** (those who understand the format best – Imaging Services)
- **Analysis, Requirements and Specifications** (Digital Preservation, though some documents may originate from other departments).
- **Quality Assurance/Plan Approval** (Digital Preservation/Imaging Services)
- **Systems Conformance/Technical Infrastructure** (Library Technology Services)
- **Content Ownership** (curators or collection managers, involvement is generally just to be informed of major decisions)

Analysis: Distill information from format/tool specifications and files within the Digital Repository Service (DRS) to build test criteria

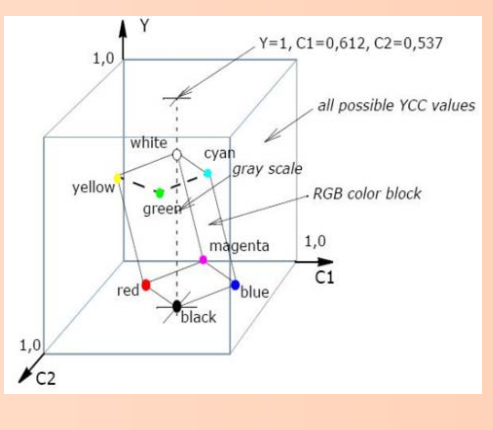
DRS – Content Groupings
Create essential "buckets" of content based on shared technical characteristics and object structures





Richard Ree/Harvard Daguerreotypes Collections
Unspecified Film Term/4050 E-6 35mm Film Term
615 files
4.723 MB each
2.838 GB total
PCD Archival Master (Cropped, YCC)
Ree in DRS as Archival Masters (177), no Producer Daguerreotypes as Production Masters (438), Boston Photo
615 files
14.949 MB each
8.97 GB total
TIFF Production Master (Cropped, RGB)
1151files
Deliverables: ~140 KB each
Thumbnails: ~5 KB each
68.154 MB total
JPEG Deliverable
JPEG Deliverable (Thumbnail)
2 JPEGs each (except Law School and Schlesinger), 129 marked with R access flag

Specifications – Significant Properties



YCC color space

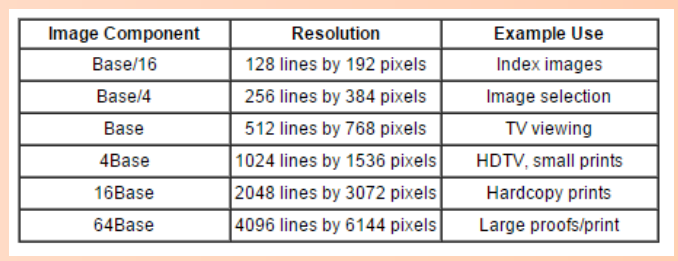
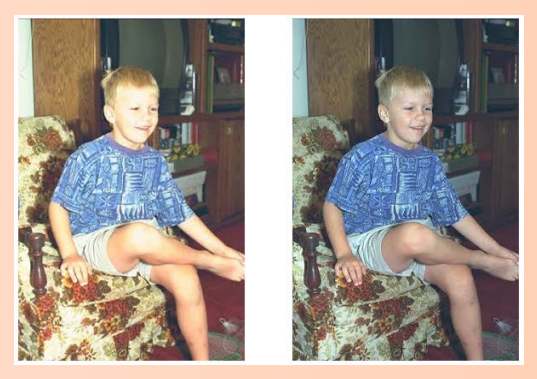


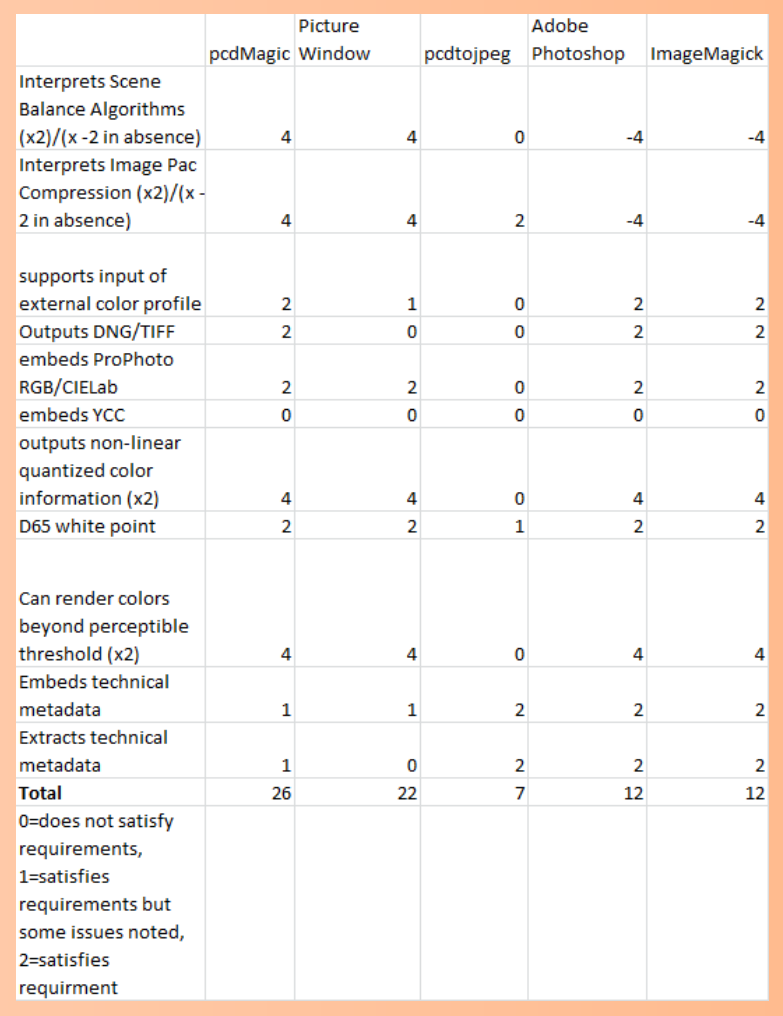
Image Component	Resolution	Example Use
Base/16	128 lines by 192 pixels	Index images
Base/4	256 lines by 384 pixels	Image selection
Base	512 lines by 768 pixels	TV viewing
4Base	1024 lines by 1536 pixels	HDTV, small prints
16Base	2048 lines by 3072 pixels	Hardcopy prints
64Base	4096 lines by 6144 pixels	Large proofs/print

Image Pac Compression



Scene Balance Algorithms for adjustments in light/color

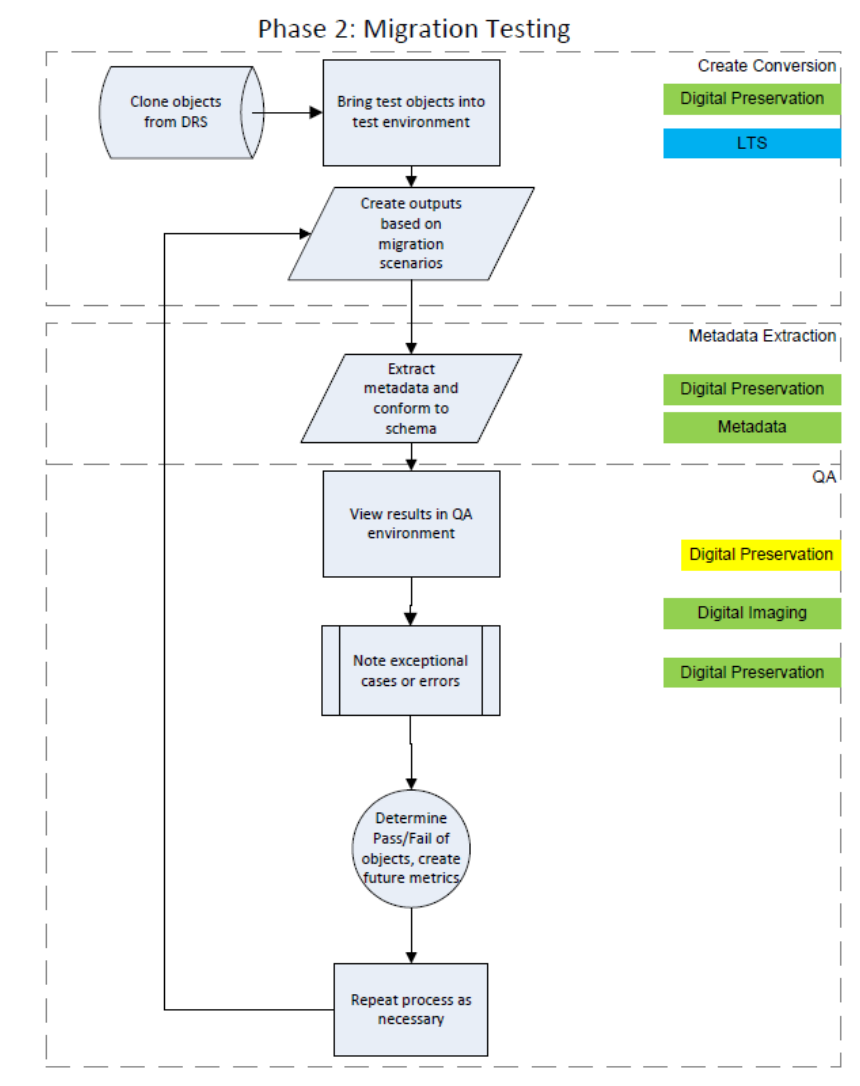
Assessment of tools
Scoring acceptability of available conversion tools based on significant properties



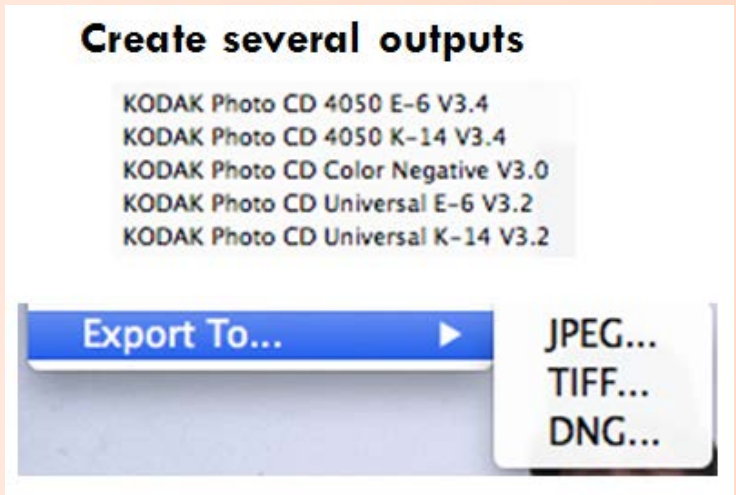
	pcdMagic	Picture Window	pcdtojpeg	Adobe Photoshop	ImageMagick
Interprets Scene Balance Algorithms (x2)/(x -2 in absence)	4	4	0	-4	-4
Interprets Image Pac Compression (x2)/(x - 2 in absence)	4	4	2	-4	-4
supports input of external color profile	2	1	0	2	2
Outputs DNG/TIFF	2	0	0	2	2
embeds ProPhoto RGB/CIELab	2	2	0	2	2
embeds YCC	0	0	0	0	0
outputs non-linear quantized color information (x2)	4	4	0	4	4
D65 white point	2	2	1	2	2
Can render colors beyond perceptible threshold (x2)	4	4	0	4	4
Embeds technical metadata	1	1	2	2	2
Extracts technical metadata	1	0	2	2	2
Total	26	22	7	12	12
0=does not satisfy requirements, 1=satisfies requirements but some issues noted, 2=satisfies requirment					

Phase 2: Testing

Component	Activity
Create Conversion	Copy sample files from DRS based on defined content buckets, perform conversions with a variety of custom settings and target formats/wrappers/codecs.
Metadata Extraction	Extract metadata from both source file and target file, compare outputs and determine significant attributes to be captured. Compare with FITS output to determine that the tool can satisfactorily parse the file.
QA	Perform quality assurance with format experts on converted files based on defined metrics



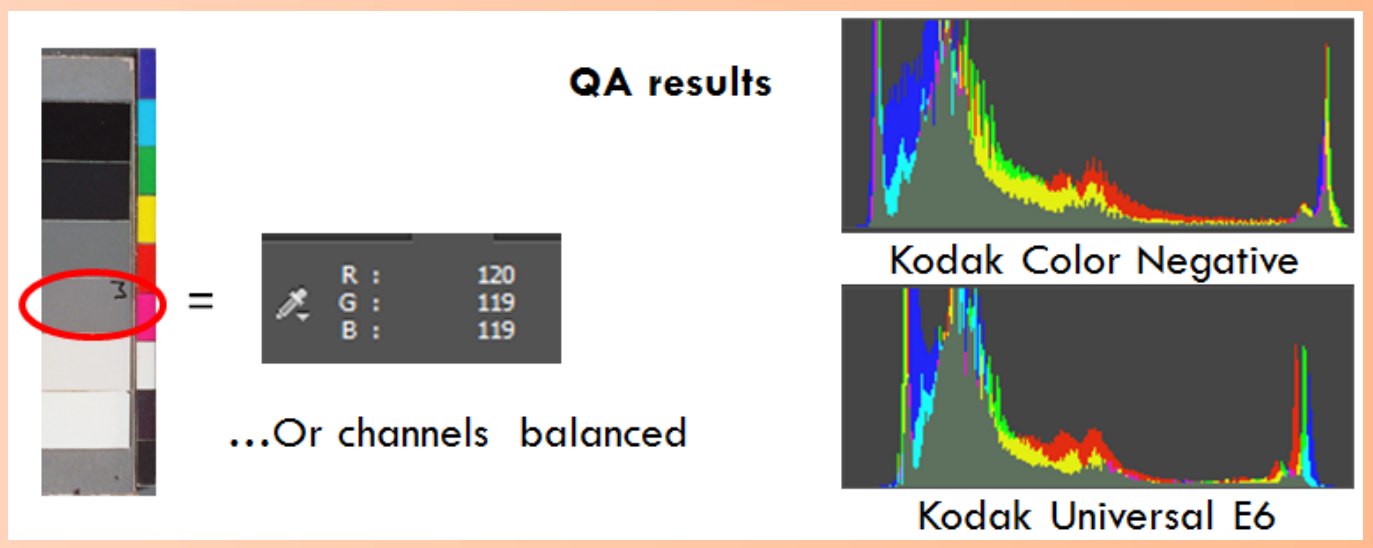
Acquire PhotoCD-specific film terms, import into pcdMagic, generate several different file format outputs for comparison.



Create several outputs

KODAK Photo CD 4050 E-6 V3.4
KODAK Photo CD 4050 K-14 V3.4
KODAK Photo CD Color Negative V3.0
KODAK Photo CD Universal E-6 V3.2
KODAK Photo CD Universal K-14 V3.2

Use Photoshop to check color balance (if image contains color bars). Compare gamut of histograms to find best range.



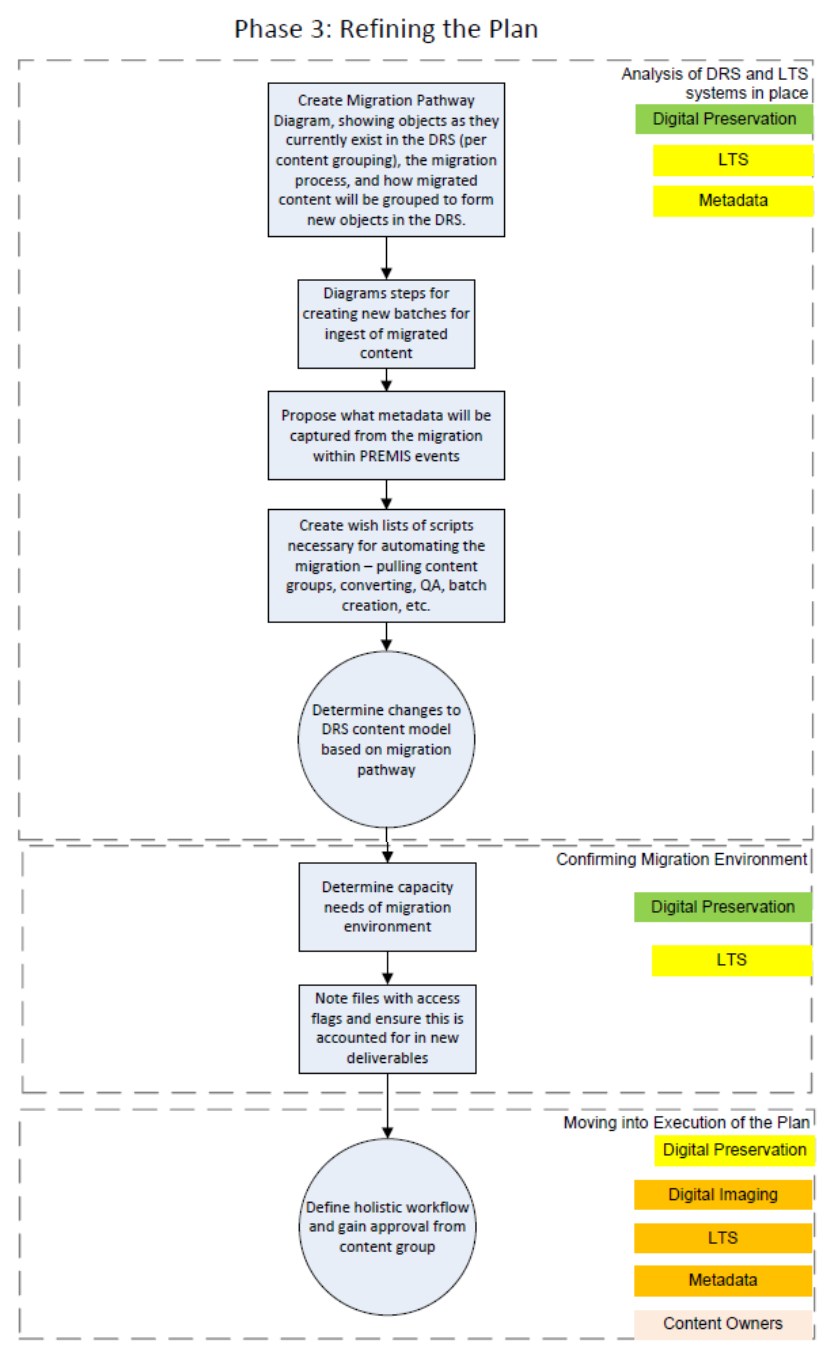
Noting subtle differences – though seemingly similar, different film terms produce slightly different results.

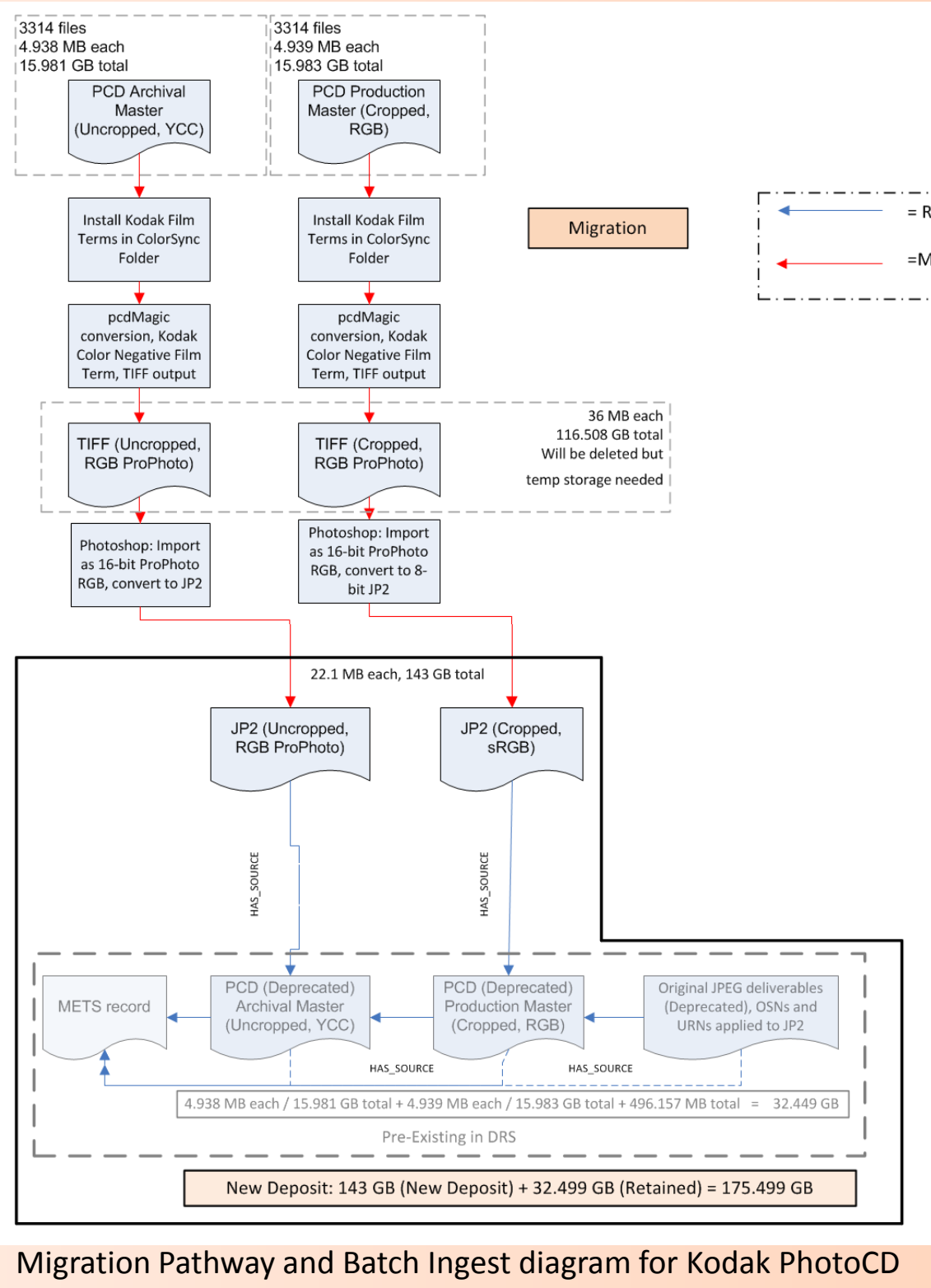
Kodak 4050 E-6 Color Profile
Kodak 4050 K-14 Color Profile

Phase 3: Refining the Plan

Component	Activity
Moving into Executing the Plan	Based on test, define migration pathway with confirmed custom settings relative to content groupings. Diagram how migrated content will be added to objects.
	Define project roles based on the Migration Workflow Diagram and RACI model assignments. Circulate this to stakeholder groups and gain approval.
Analysis of DRS and LTS systems in place	Diagram modifications to content model, file-to-file relationships, and ingest processes in order to accommodate migrated content.
	Create and approve a plan with Library Technology Services for batch ingests, design method to pull files from DRS based on content groupings and bring into migration environment.

Component	Activity
Confirming Migration Environment	In confirming that the technical environment and processes for connecting them are in place, make sure the following questions are answered: • What is the environment for the migration? What is the total filesize of everything that will be pulled, staged, generated, and re-packaged? • What services are needed to transfer material and will any tools need to be built for this process? • Does the chosen tool interface well with automated scripts such that migration can be performed based on designated content groups? • How will QA be performed en masse? • How will new files and their relationships be handled through batch ingest processes? • Are there other external dependencies which are required to render the file (streaming service, application, etc.) or for it to function as an object within the DRS? • Do any exceptions need to be made in separating content based on the Access Flag? This will be especially relevant for deliverable content.





3314 files
4.938 MB each
15.981 GB total
PCD Archival Master (Uncropped, YCC)
3314 files
4.939 MB each
15.983 GB total
PCD Production Master (Cropped, RGB)
Migration
= Relationship
=Migration Pathway
Install Kodak Film Terms in ColorSync Folder
Install Kodak Film Terms in ColorSync Folder
pcdMagic conversion, Kodak Color Negative Film Term, TIFF output
pcdMagic conversion, Kodak Color Negative Film Term, TIFF output
TIFF (Uncropped, RGB ProPhoto)
TIFF (Cropped, RGB ProPhoto)
36 MB each
116.508 GB total
Will be deleted but temp storage needed
Photoshop: Import as 16-bit ProPhoto RGB, convert to JP2
Photoshop: Import as 16-bit ProPhoto RGB, convert to 8-bit JP2
22.1 MB each, 143 GB total
JP2 (Uncropped, RGB ProPhoto)
JP2 (Cropped, sRGB)
HAS_SOURCE
HAS_SOURCE
METS record
PCD (Deprecated) Archival Master (Uncropped, YCC)
PCD (Deprecated) Production Master (Cropped, RGB)
Original JPEG deliverables (Deprecated), OSNs and URNs applied to JP2
HAS_SOURCE
HAS_SOURCE
4.938 MB each / 15.981 GB total + 4.939 MB each / 15.983 GB total + 496.157 MB total = 32.449 GB
Pre-Existing in DRS
New Deposit: 143 GB (New Deposit) + 32.499 GB (Retained) = 175.499 GB

Migration Pathway and Batch Ingest diagram for Kodak PhotoCD

Diagram migration pathway of the content, paying particular attention to relationships between files (masters, derivatives, related metadata files) and object structures within the repository:

-How content looks now

-What happens to the content at each step of the migration

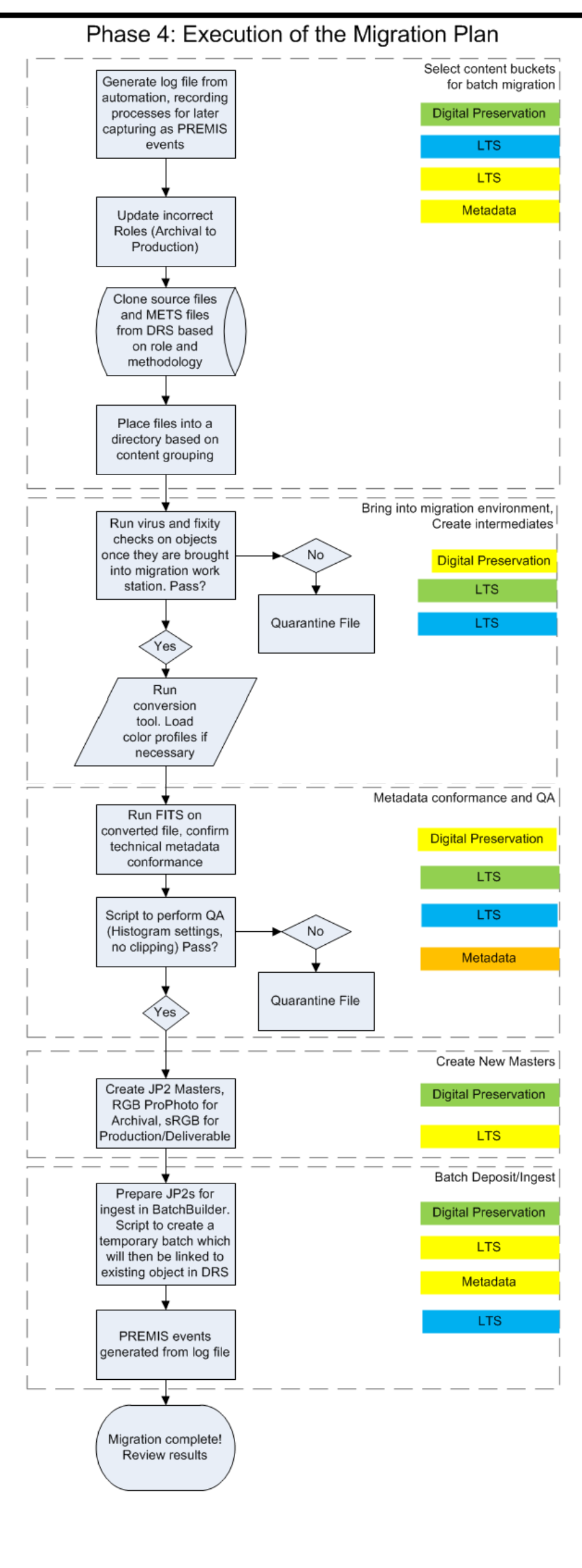
-How content will be structured such that it complies with repository ingest policies

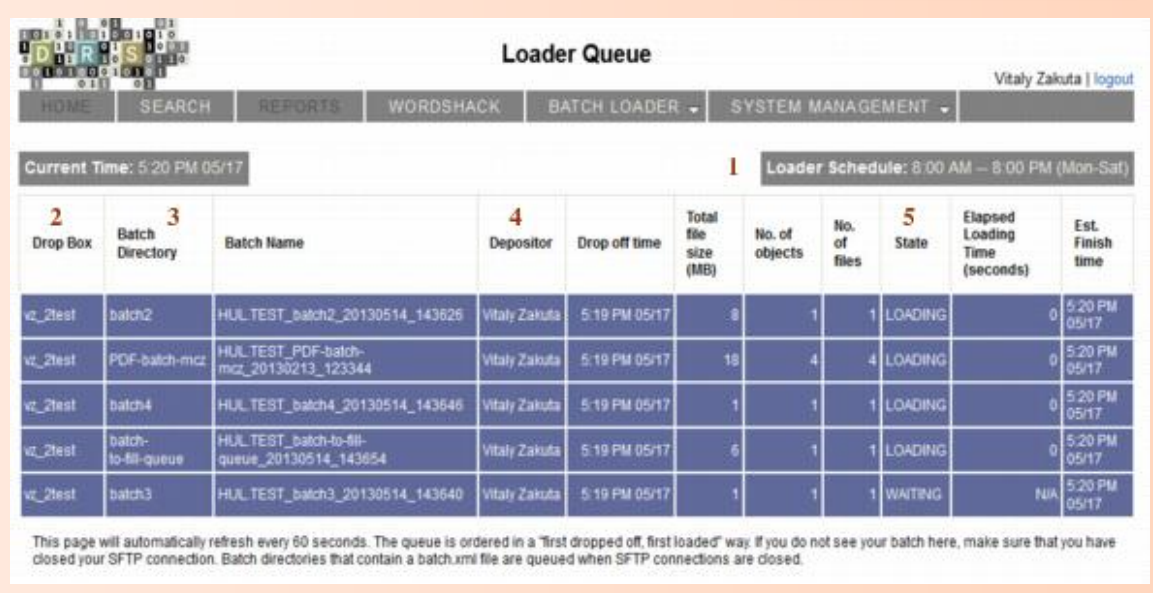
Begin to think about what tools and services are needed to automate this process.

Phase 4: Executing the Plan

Component	Activity
Schedule Migration	Set timelines relative to existing projects and departmental workloads. Anticipate setbacks that could occur as well as deadlines that need to be met.
Custom development	Develop scripts for automating the process Prepare Migration Environment • Locate grouping of objects based on DRS criteria (e.g. Production Masters in a given format, methodology statements, relationship to a list METS record IDs, etc.) • Move grouping of objects to migration environment (e.g. async), mark status of object to avoid offline issues in DRS • Prepare log file to record migration events, later for recording within PREMIS • Determine process for recording file's relationship to its object such that new content can be joined with the object during ingest Perform Migration • Interface with conversion tool to convert files based on custom settings and outputs

Component	Activity
Custom development	QA • QA results based on pass/fail metrics (e.g. no color clipping in Photoshop, audio file successfully plays in multiple browsers through SDS, etc.) Ingest • Create batch submissions with a particular focus on generating new metadata (from migration), creating new relationships with files in the DRS ('HAS_SOURCE', Deprecating replaced file), and running file fixity and characterization validation. Reassign URNs for replacing deliverables (or create new ones), may also need to deactivate current URNs. Also ensure that new deliverables match the access flags of the deliverables that they are replacing. Create batch deposit with batch control files designed for content buckets relative to their intended relationships structure within a content model





Current Time: 5:20 PM 05/17
Loader Schedule: 8:00 AM – 8:00 PM (Mon-Sat)

Drop Box	Batch Directory	Batch Name	Depositor	Drop off time	Total file size (MB)	No. of objects	No. of files	State	Elapsed Loading Time (seconds)	Est. Finish time
vz_2test	batch2	HUL.TEST_batch2_20130514_143626	Vitaly Zakuta	5:19 PM 05/17	8	1	1	LOADING	0	5:20 PM 05/17
vz_2test	PDF-batch-mcz	HUL.TEST_PDF-batch-mcz_20130213_123344	Vitaly Zakuta	5:19 PM 05/17	18	4	4	LOADING	0	5:20 PM 05/17
vz_2test	batch4	HUL.TEST_batch4_20130514_143646	Vitaly Zakuta	5:19 PM 05/17	1	1	1	LOADING	0	5:20 PM 05/17
vz_2test	batch-to-fill-queue	HUL.TEST_batch-to-fill-queue_20130514_143654	Vitaly Zakuta	5:19 PM 05/17	6	1	1	LOADING	0	5:20 PM 05/17
vz_2test	batch3	HUL.TEST_batch3_20130514_143640	Vitaly Zakuta	5:19 PM 05/17	1	1	1	WAITING	N/A	5:20 PM 05/17

This page will automatically refresh every 60 seconds. The queue is ordered in a "first dropped off, first loaded" way. If you do not see your batch here, make sure that you have closed your SFTP connection. Batch directories that contain a batch.xml file are queued when SFTP connections are closed.

Harvard's BatchBuilder tool for managing ingests into the DRS

Necessary scripts and software augmentation to enable automatic workflow:

-Clone files from DRS based on role and methodology, METS file to follow the file to later link with existing objects upon ingest.

-Batch convert in **pcdMagic** based on content groupings

-Run **FITS** on converted files to confirm technical metadata conformance

-**Photoshop** API scripting for QA (histogram checking) and conversion to target format (JP2, ProPhoto RGB for Archival Masters, sRGB for Production

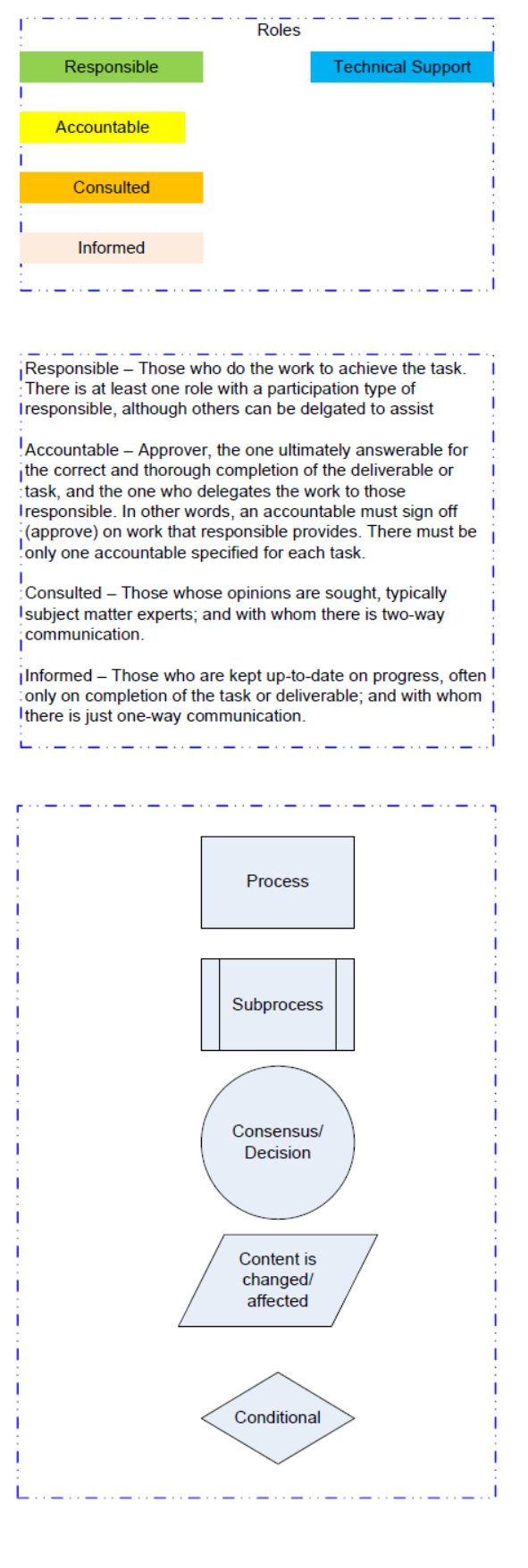
-Devise strategy within **BatchBuilder** (ingest processing software) to create temporary batch ingest which is later linked with existing objects

-Reassign URNs, OSNs, and Access Flag from old deliverables to new, change status of old masters to "Deprecated"

Phase 5: Project Wrap-Up

Component	Activity
Verify results Post-Ingest	Verify that migrations were successful (batch QA reports, ingest reports)
Encapsulate project Deliverables/ Artifacts	Ensure that all deliverables/artifacts and project documentation is complete. Determine long-term disposition of the documents (what is deposited in DRS and what is kept in separate project folders in share drive). Confirm plan to deaccession any content.

Workflow Legend



HARVARD
LIBRARY

NDSR BOSTON

Acknowledgements

I would love to thank everyone at Harvard Library whose insights contributed greatly to the development of the framework. In particular I would like to thank Andrea Goethals, Manager of Digital Preservation and Repository Services, who supervised me during this project and without whom creation and refinement of the plan would not have been possible. I would also like to thank my contacts in Digital Imaging and Library Technology Services for lending their time, resources, and expertise.

Joey Heinen
National Digital Stewardship Resident
Harvard Library
joeygheinen.jh@gmail.com
917-684-9495