Developing a Framework for File Format Migrations

Joey Heinen
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1 (617) 373-3669
j.heinen@neu.edu

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1 (617) 495-3724
andrea_goethals@harvard.edu

ABSTRACT

In this paper, we describe the development of a file format migrations framework at Harvard Library, using one migration case study, Kodak PhotoCD images, to demonstrate implementation of the framework.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Frameworks for digital preservation; Preservation strategies and workflows; Innovative practice;

Keywords

Format Migrations; Migration Frameworks; Obsolete Formats

1. INTRODUCTION

As is well known to memory institutions, the act of preservation is never done, particularly once an object has been digitized. Digital material is just as susceptible to obsolescence as analog formats. There are a number of digital preservation strategies that can be employed in order to protect the usefulness of data, for example emulation, normalization and migration. Migration is chosen as a digital preservation strategy when the aim is to move content from its previously tenuous origins to a format with much greater promise in terms of support and usage [1]. Harvard Library uses migration as a primary preservation strategy because the Library's goal is to continue to provide networked access to digital collections on emerging platforms, without requiring researchers to physically come to the Library or to install special software.

Many institutions have demonstrated successful digital format migration projects (largely text-based) which focused on identification and preservation of significant properties within the format. However, few examples exist for how these projects can scale to a larger framework that can be continuously adapted for future format migrations, and for thousands or millions of files. At Harvard Library as a National Digital Stewardship Residency [2] project, one such possible framework was created. In order to test the viability of this generic framework, the project included the development of migration plans for three obsolete formats within Harvard Library's Digital Repository Service (DRS) [3] – Kodak PhotoCD, SMIL playlists, and RealAudio. While each format has its own challenges that will introduce deviations to a workflow, there are certain processes that will always be included in the migration workflow and plan. This paper does not diagram all aspects of the framework but outlines the main phases and components for creating a migration plan for a format. More information and documentation can be obtained by contacting Harvard Library directly.

Kodak PhotoCD is one of the first formats for digitized analog photographs and was used at Harvard largely for early photography and daguerreotypes collections. Real Audio and SMIL playlists were used for audio delivery at Harvard Library. These older formats are no longer deposited to the DRS but there is a great deal of content in the DRS in these legacy formats that needs to be migrated to modern formats. This project was made possible through the National Digital Stewardship Residency which allowed the resident (Joey Heinen) to develop this project over the course of nine months. Due to the time constraints of the project, the SMIL playlist project was only planned at a high level. The plan for the Kodak PhotoCD migration is complete though Harvard has yet to perform the migration.

2. CONTEXT FOR THE FRAMEWORK

The Harvard Formats Migration Framework is intended to inform the migration of obsolete formats regardless of the file format. While the specific format will necessitate variations to the overall framework, the framework will depict the general processes that must occur for each format in relatively the same sequence.

While the hope is that the framework can inform migration projects at large regardless of institutional context, there is obviously quite a bit that inevitably must be Harvard-specific – in particular, Harvard's organizational structure, policies, and digital preservation repository. Digital Preservation as a department resides within Preservation Services though also maintains strong ties with Library Technology Services (LTS), library-dedicated IT staff. Responsibilities for the Digital Repository Service (DRS) are shared between these departments with Preservation Services serving as the business owner and LTS as the technology owner. The DRS is both a preservation and an access repository. It provides Harvard affiliated owners with a set of professionally managed services to ensure the usability of securely stored digital objects over time.

There are a few DRS concepts that must be understood to understand the migration framework:

- A DRS **object** is a coherent set of content that is considered a single intellectual unit for purposes of description, use and/or management: for example a particular book, web harvest, serial or photograph.
- Each object conforms to a single **content model** which defines the object type (audio, still image, etc.). Content models define the supported file formats, object

structure, file and object relationships, roles and other key metadata.

- As defined by PREMIS "a **file** is a named and ordered sequence of bytes that is known by an operating system. A file can be zero or more bytes and has a file format, access permissions, and file system characterizations such as file size and last modification date." [4]

In a format migration plan, the files are the source for the migration and the plan will need to consider how to add migrated files to an existing object (which can contain different generations of the same files).

The associated content model of a format will also greatly affect the resulting migration plan. The content model affects which source file will be selected for the migration (the highest-quality version when possible) as well as how newly-migrated content is added to the repository, including how new relationships will be formed with the existing content. More complex formats may work interdependently with other files in order to produce the final results, such as the SMIL playlists which assist in delivering the RealAudio files. Understanding the content model, in particular the relationships that must be maintained or modified across migrations, is a crucial part of developing the migration plan.

2.1 Example – Kodak PhotoCD

Harvard Library preserves more than 7000 PhotoCD (PCD) files within the DRS and due to increasing difficulty in accessing (and thus preserving) these files over time, Digital Preservation Services decided to embark on a project to migrate these files to a modern target format. So that the Library would have a blueprint for conducting migrations for future obsolete formats, the PCD migration was used as a test case for developing the generic migration framework, noting the processes that must occur and generally in what order. While the overall generic framework had largely been designed as part of this project before testing it on PCD, it became an iterative process, updating the generic framework as experience was gained with this actual test case.

2.2 Related Projects

The project began with a literature review of migration projects. While an example could not be found of a format-agnostic migration framework that had been put to into production within an institution, many projects proved inspirational to the development of this framework, especially in the early stages. Workflow designs and models for building a migration plan from start to finish exist in the form of single projects/formats (National Library of New Zealand's WordStar to HTML4 [5], the Austrian National Library's TIFF to JP2 [6]) as well as larger institutional models for depicting roles and responsibilities (see Acknowledgements). Other projects demonstrated use of integrated tools to characterize/validate, convert, and QC migrated content (Austrian National Library). Others discussed use of registries and knowledge-bases to contain data on recommended tools and platforms for format migration (University of Illinois at Urbana-Champaign's Digital Preservation Interoperability Framework/Conversion Software Registry [7]) or to design and enforce holistic workflows and policies on migration (Technical University Vienna/AARIT's Plato [8], Norwegian University of Science and Technology's Multi-Criteria Decision Making model [9]). While these projects were not directly referenced in the design of the Harvard framework, the review helped to identify shared ideologies in what constitutes a successful migration and how to connect systems and technologies with theoretical processes.

3. THE FRAMEWORK

The specifics of this framework are much too large to describe here in detail, but the main components are stakeholder identification, migration workflow, and migration deliverables.

3.1 Stakeholder Identification

The identification of stakeholders first is deliberate – without clear roles and responsibilities, the migration project cannot start. Depending on the type of content, the particular departments and individuals may vary but the roles involved will stay somewhat consistent. The framework includes the following key stakeholder roles:

1. **Project Management** (those managing the overall migration project)
2. **Format Experts** (those who understand the format best)
3. **Content and Metadata Analysts** (those analyzing the content and metadata to be migrated and creating requirements documents and specifications)
4. **Plan Reviewers** (those reviewing plans and specifications)
5. **Systems and Technology Experts** (those helping to design the technical workflow and providing development and infrastructure support for the migration)

- **Content Owners** (curatorial stewards of the content to be migrated)

At Harvard, for format migrations, Digital Preservation Services plays the Project Management role, and serves as the primary Content and Metadata Analysts. The Format Experts vary, for the Kodak PhotoCD migration it is Imaging Services; for the SMIL playlists and RealAudio files it is Media Preservation Services. The Plan Reviewers include a variety of people across departments, and the Systems and Technology Experts role is played by Library Technology Services. The Content Owners vary depending on the content to be migrated, but will generally come from Harvard libraries, archives or museums.

3.2 Migration Workflow

The migration workflow can be broken down into five phases:

1. Plan for Test
2. Test
3. Refine Plan
4. Execute Plan
5. Verify Results and Wrap-Up Project

The workflow includes the creation of the migration plan as well as the actual migration. Each project phase can be further broken down into sub-phases and activities that may or may not produce deliverables.

- **Workflow Phases** are the five high-level parts of the migration workflow, each of which is further broken down into **Workflow Sub-phases** containing **Workflow Activities** (actions common to any migration)
- **Deliverables** include the migrated content itself, documentation or metadata generated along the way, diagrams, plans, or new revelations in digital preservation policies (e.g. storage and retention plans).

3.3 Migration Deliverables

The framework defines a set of deliverables for each phase, described here.

Phase 1: Plan for Test

- **Stakeholder Chart:** Identifies the departments and/or staff members who will fill roles during the migration project.
- **Format Specifications:** Where possible acquire authoritative descriptions of the relevant formats (formats to migrate but possibly also for the formats that will be migrated to)
- **Format Analysis Report:** Conclusions drawn from format technical specifications to determine significant properties, target formats, and possible conversion tools. Also include conclusions drawn from DRS metadata (or other relevant Harvard-specific sources).
- **Content Grouping Diagram:** Make-up of migration source files, their relationships to other files within an object, and the noteworthy technical attributes that will distinguish the ways that they are migrated (e.g. methodology, role, owner code, etc.). Includes useful SQL as an appendix
- **Target Formats/Conversion Tool Analysis:** Conclusions on target formats and conversion tools will be used in the test phase (and which ones will not), a scoring template which rates a tool/format's compliance with the defined significant properties of the format.
- **Migration Environment Specifications:** A list of necessary tools, plug-ins, and other software-based needs and the necessary OS/platforms/processors needed to support the software. Consider short-term storage capacity needs if necessary.

Phase 2: Test

- **Testing Conclusions Report**: Findings of the tests, the testing parameters, metrics for determining acceptability of the conversion, analysis of embedded metadata, and decisions on the best courses of action for the migration.

Phase 3: Refine Plan

- **Migration Pathway Diagram:** How migration will be performed based on content sub-groups, how migrated files will be created relative to conversion tools and custom settings, target formats and how these will be deposited and related to existing files in the DRS.
- **Migration Workflow Diagram:** Workflow processes mapped against RACI model (roles broken down into Responsible, Accountable, Consulted, and Informed) for stakeholder involvement. The workflow is broken into the 5 migration phases. Within each RACI grouping, define the plan components (see Format Migration Framework section). Uses shapes to correspond with the action (Process, Sub-process, Consensus/Decision, Changes to Content, Conditional Factors).
- **Migration Plan:** This is a comprehensive summary of all conclusions drawn from analysis and testing. Emphasis will be placed on necessary tools and systems for grouping, converting, and ingesting files based on content groupings.
- **Metadata Mock-up:** A wishlist for augmentation metadata to include information about the migration (processes, tools, etc.), generally for recording migration-specific PREMIS events.
- **Batch Ingest Mock-up:** A step-by-step process of how batches will be created based on migrated content grouped along with existing files within an object.

Phase 4: Execute Plan

- **Migration Checklist:** Record of the migration process, including key staff involved and tools used

Phase 5: Verify Results and Wrap-Up Project

- **QC Report:** Record of the verification of the converted files (passes based on decided metrics through QC tools if available).
- **Migration Conclusions:** Summarize lessons learned noting any anomalies or adjustments made along the way that might help to inform modification of framework or plan documentation.

3.4 Migration Workflow Example: Kodak PhotoCD Images

In this section, each phase and sub-phase of the generic framework is briefly described and then followed by an illustration from developing the PCD migration plan.

3.4.1 Phase 1: Plan for Test

3.4.1.1 Sub-Phase 1: Project Start-Up

Project start-up involves identifying key stakeholder roles and responsibilities, and setting the stage for the analysis and planning which must imminently take place. Acquiring technical specifications and format reports, exploring the provenance of the format within the institution/collections, and securing a technical environment for performing the basic analysis are all essential first steps. Additionally, it is essential to identify parallel library projects that will affect the migration from the outset so that they are embedded within the plan.

For the PCD plan, staff in Imaging Services served as Format Experts and participated with others as Plan Reviewers. They helped to analyze the significant properties of the format and to design the testing environment. Library Technology Services would be responsible for Systems and Technology Experts. Digital Preservation would be responsible for Project Management and Content and Metadata Analysts.

A project to migrate all metadata from an older version of the DRS to a newer version was running concurrently to the development of the PCD plan. The metadata migration project made changes to the content model associated with PCD images (most importantly in how file-to-file relationships are described). This metadata migration project was considered at many steps of the PCD migration plan.

3.4.1.2 Sub-Phase 2: Analysis

The first steps of the analysis sub-phase are to research the format specifications to identify the important technical characteristics of the format and to analyze the DRS metadata to break content down into groups relevant to the migration. This analysis should result in an early understanding of what the requirements might be for target formats and tools (for conversion, metadata extraction, and so on). Naturally, the technical characteristics and ways that content can be broken down will vary considerably based on the format, but this generic component will always be a necessary precursor to developing a format migration plan.

Kodak PhotoCD is a proprietary format that was popular in the late 1990s as a means of creating digital surrogates of analog photographs and slides. While it was at first adopted as an archival format it was eventually noted that its use of proprietary rendering software and applications as well as its unique forms of compression and color encoding were contributing factors in the format's eventual obsolescence (broadly discussed as early as 2005) [10].

Based on analysis of various technical specification documents, web forums, and white papers, the following significant properties were defined for the PCD format:

- PCD used PhotoYCC, a unique color space for segmenting luminance information from two chrominance channels (third channel interpolated) and to encode color information that goes beyond that which is conventionally contained within 255 decimal values. The color space is device independent meaning that it is designed to be rendered on any number of output devices (both analog and digital) [11]. Few file formats can support and render this color space.
- The PCD format supports a number of Scene Balance Algorithms (SBAs) (used for automatic lightness and color-balance adjustment) that can be applied at the time of scanning. This means that for different photo stocks and materials, different algorithms could be used to encode the scanning data to account for nuances in light and color [12]. SBAs need to be understood and accounted for by a conversion tool so that color (chrominance) and light (luminance) are presented as accurately as possible.
- In addition to technical metadata, provenance metadata from the digitization and from the disk encoding process history can be found embedded within the file.
- Image Pac compression, used by the PCD format, is a very efficient form of mathematically lossless encoding which may not be compatible with reporting or conversion tools. For example, ImageMagick [13] reads the Harvard PCD files as 768 x 512 (the Base image rather than Base16) and reports the compression scheme as "undefined." It will be important that the migration tools know how to unpack the images and read them at their fully uncompressed resolution (2048 x 3072) [14].

The DRS metadata is stored in an Oracle database. For DRS analysis, the database needed to be queried using SQL in order to explore the metadata looking for key technical and historical differences among the content as well as the relationships between content in this format and other formats. The results of these queries and analysis of the data is expressed pictorially in the **Content Grouping Diagram.** The most useful metadata for classifying the PCD files into groups was found in the methodology field, which is where free-text narratives described the digitization process for the file. This metadata was used to group the files into three essential groups based on their collections – The Harvard Daguerreotypes, the Horblit Collection, and the Richard H. Ree Collections. The first two collections feature early photography holdings (mostly daguerreotypes) that were some of the first photo digitization projects at Harvard in the late 1990s. They both employed the Kodak PhotoCD scanning process though used a unique Scene Balance Algorithm to account for different photo stocks that were used to initially photograph the images objects. The process used to create the Ree Collection is a little less clear cut, especially given this line from the methodology statement associated with this content:

"Ree's PhotoCD format images were processed using Adobe Photoshop 6.xx and 7.xx. The PhotoCD files were imported into Photoshop as 16 bit RGB TIFF files using the built-in import module with the "universal E-6" film term. Each image was individually processed to compensate for any obvious color casts and to achieve, to the extent possible, natural tone and color."

It is not noted how the images were digitized, simply that they were imported as digital. It also seems that images were individually corrected at the discretion of the Imaging Technician such that a monolithic film term setting wouldn't help to account for any of the original color or light settings (even if "universal E-6" was used to import the images into PhotoShop). Unfortunately no other provenance documentation exists from the original digitization or deposit of the digital images so the best that can be done is to analyze additional metadata within the DRS (and also to keep a sharper eye on this collection during conversion testing).

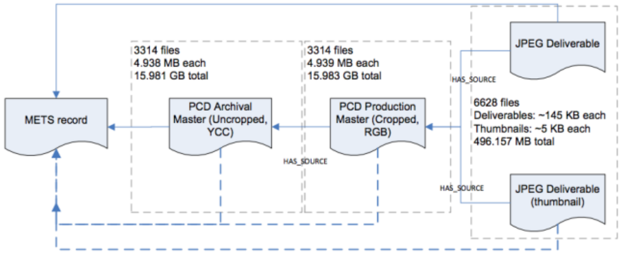


Figure 1: Content Grouping Diagram for a Still Image object from the Horblit Collection. This particular grouping shows PCD as both an Archival Master (Uncropped) and Production Master (Cropped) which will both be used as migration sources.

Additional DRS metadata was helpful for designing the **Migration Pathway Diagram** for each set of files that could be migrated as a group. "Roles" metadata defines the file's placement within the production workflow, namely for the Still Image content model if the file is an Archival Master, Production Master, or Deliverable. This was useful for determining which file to use as the source for the migration. For the Horblit Collection, PCD was used for both Archival and Production Masters with JPEGs as deliverables. The Archival Masters were fully uncropped including color bars for calibrating the scanning equipment to the imaging environment. The Production Masters were cropped and used to generate deliverable JPEGs for the web. It was decided that new Archival and Production Master images would need to be generated during migration. For the Harvard Daguerreotypes and Richard Ree collections, a PCD Archival Master (cropped) was used to generate a TIFF Production Master, which was used as the source for generating JPEG deliverables. It was decided that the TIFFs would be removed since they were generated using inferior PCD conversion software that did not account for SBA settings. For this case the PCD Archival Masters would be used to generate both a JP2 Archival and Production Master.

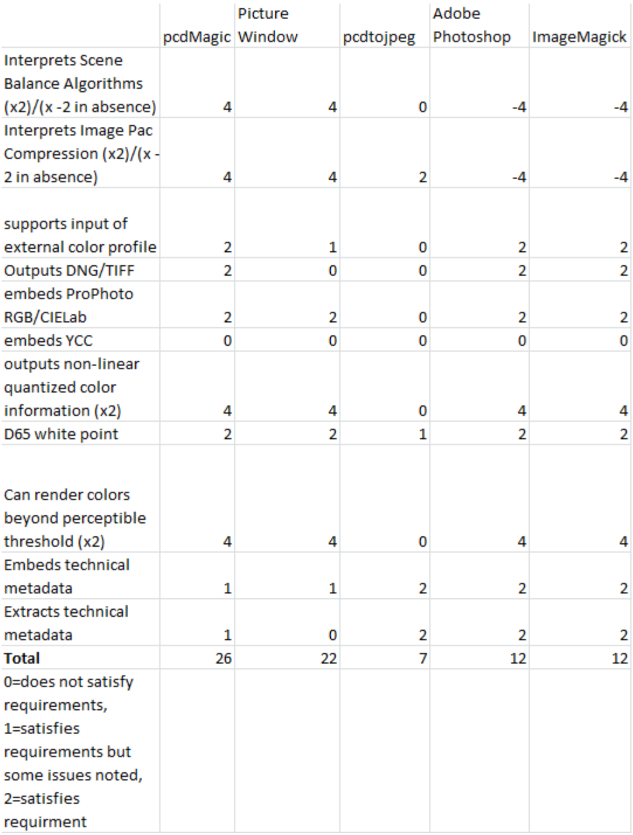
Other metadata was also useful in building the overall framework and migration plan, but in unexpected ways. The analysis uncovered errors in manually-submitted metadata for some files, specifically for metadata about Color Space, Compression, Dimensions, Scanning Systems, Vendor/Producers, and Roles. For example, the Production Masters in the Horblit Collection were all listed incorrectly as RGB Lossless images instead of YCC Image Pac. This would need to be corrected before creating the final batches for DRS ingest. Additionally, all the images from the Harvard Daguerreotypes that should have been listed as Archival Masters were marked as Production Masters, metadata that would also need to be corrected before DRS ingest (or even before execution of the migration in the event that scripts are used to pull PCD images from the DRS based on their "Role").

3.4.1.3 Sub-Phase 3: Confirming Migration Criteria

This sub-phase includes working closely with the Format Experts to confirm the analysis results - that the format's significant properties have been defined, the results of the metadata analysis look correct, and that there is clear criteria for choosing among conversion tools.

In the PCD case, the ideal target formats which emerged out of the analysis phase were confirmed by Imaging Services staff. They confirmed that the YCC color space is not supported by many image formats but can be mapped to ProPhoto RGB with minimal-to-no loss in information [15], CIELab also demonstrating good results [16]. Of course, all of this would be inconsequential if there were no available tools for performing these conversions, leading into an analysis of the available tools.

A scoring system as shown in Figure 2 was used to compare conversion tools as well as for choosing possible target formats (not pictured), resulting in the **Target Formats/Conversion Tools Analysis** report. A score was applied to each of the criteria (based on the defined significant properties of the format) which for some criteria involved a weighted score. In some instances an especially important criteria could incur a negative fee if the tool did not support this feature, meaning that the tool in general was not sufficient for use in the migration. In scoring tools based on their ability to meet the needs of the format migration and adding up the scores to generate a final value, it was much clearer to see which tools would generate a more desirable outcome, and especially which tools were unacceptable and would not require inclusion in the actual migration testing.



	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					

Figure 2: Scoring acceptability of conversion tools

Based on this analysis Digital Preservation and Imaging Services decided that pcdMagic was the best conversion tool based on its ability to meet all of the essential criteria including its ability to interpret YCC, SBAs, and Image Pac compression. It could output both TIFF and DNG with a ProPhoto RGB color space. Additionally, it can accept external color profiles for more precise rendering of the color and light information (as an alternative to SBAs). Fortunately, Imaging Services owned color profiles that were specific to film terms used in some of the original PCD scanning software, something that would prove to be a boon to a successful migration plan.

3.4.1.4 Sub-Phase 4: Metadata Analysis

This sub-phase is an exploration of the tools that can best best characterize the format and/or provide process history information about the conversion process.

Though pcdtojpeg was found to not be an ideal tool for converting the format, it was the best at outputting provenance metadata about the file (scanning information, SBA settings, etc.). ImageMagick, another rejected conversion tool, was also a good metadata extraction tool because it was able to extract Exif metadata and technical metadata about the RGB channels. Exiftool was also used for metadata analysis, particularly for analyzing the images post-conversion. In the Exiftool output, the DNG files produced from the PCD files would present a color space of "pcdMagic DNG Profile" under "Profile Name" whereas the TIFF files would present a color space of "ProPhoto RGB" under "Profile Description." This led to a decision to choose TIFF as an intermediate output during the conversion because the color space is more standardized and a better choice for preservation.

3.4.1.5 Sub-Phase 5: Moving Into Test Phase

At this point in the workflow the tools and target formats for the conversion have been decided; this sub-phase includes additional testing to determine some of the conversion details including how the tools would be run and any tool parameters.

For the PCD plan, this mostly came down to the Scene Balance Algorithms and how to most accurately depict and capture color information from the image. The environment for perfoming the migration was determined, which in this case was a Mac OS X environment (the most recent release of pcdMagic works with the OS X environment and had no additional dependencies besides an optional addition of color profiles in the ColorSync folder). pcdMagic is available for both Mac and Windows platforms however the Mac version is the only version that allows for external color profiles. For the testing phase a test laptop was used knowing that it would be possible to transfer the license for the tool to a production workstation when ready to move to Phase 4: Execute Plan.

3.4.2 Phase 2: Test

3.4.2.1 Sub-Phase 1: Create Sample Conversions

In this sub-phase a representative subset of the content is converted in preparation for analyzing the results together with the Format Experts.

For each PCD content grouping (determined by the methodology/collection with which the image is associated), 6-8 images were selected for testing. All five of the Kodak color profiles provided by Imaging Services (Color Negative, 4050 E-6, 4050 K-14, Universal E-6, Universal K-15) and a sampling of the general settings provided within the tool were tested (largely for comparison to demonstrate inadequacy of the pre-existing settings). The images were output in both TIFF and DNG format using various settings in order to determine the more ideal target format.

3.4.2.2 Sub-Phase 2: Assessment of Sample Conversions

In this sub-phase a combination of manual and automated tasks are performed with input from the Format Experts to make final decisions about how the migration will be performed and to verify that the conversion will be acceptable.

The PCD test conversions were viewed within PhotoShop. For additional comparison, multiple film terms were selected for each image in both TIFF and DNG formats. After refreshing the images, RGB histograms were consulted to make sure that no clipping of information had occurred and to see which images produced the widest gamut with the most evenly distributed waveform throughout. In some instances Imaging Services staff would make final judgments based on visual appearance, determining which images presented the best real-world results (not overcompensating in any of the RGB channels). The key decisions as documented in the **Testing Conclusions Report** were:

1. For the Horblit Collection, the Kodak Color Negative film term produced the best results. This was commensurate with the methodology statement.
2. For the Harvard Daguerreotypes, the 4050 E-6 film term produced the best results. This was commensurate with the methodology statement.
3. For the Richard H. Ree Collection the 4050 E-6 film term generally produced the best results (though in some cases was not as definitive). This is not commensurate with the methodology statements which said that the Universal E-6 film term was used though this does not appear to be the case. It will be necessary to decide if this group will require a more detailed conversion process where all 177 images are converted with their own unique settings.

During this process no discernible differences were seen between TIFF and DNG outputs (confirmed by subtracting pixel information from images and also comparing histogram readings) and that cropped and uncropped versions of the same image produced virtually identical color mappings (with the exception of borders and presence of color bars). However, as noted earlier, the color space associated with DNG was less preferable to the ProPhoto RGB found in the TIFF output.

As an extra step of quality control, characterization tools were used to ensure that embedded metadata was not lost (largely provenance).

3.4.3 Phase 3: Refine Plan

3.4.3.1 Sub-Phase 1: Analysis of Systems in Place

In this sub-phase the migration plan, which up until this time has been largely theoretical, is integrated with the Harvard Library infrastructure. Decisions need to be made about how the DRS files relate to the files that will be produced in the migration, and how the files produced in the migration will be integrated into existing DRS objects, and which files will be retained.

In order to gain insight and approval from all relevant stakeholders, the migration process is expressed pictorially in the **Migration Pathway Diagram (See Figure 4 for a PCD example).** The overall process employed for the entire PCD format migration (including initial planning and analysis phases) is expressed in the **Migration Workflow Diagrams** along with stakeholder roles and responsibilities. A narrative version is outlined in the **Migration Plan** document.

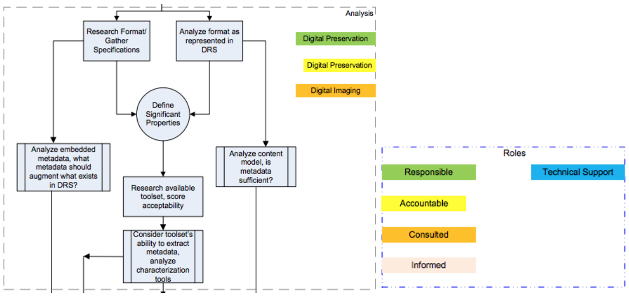

Analysis
Research Format/ Gather Specifications
Analyze format as represented in DRS
Digital Preservation
Digital Preservation
Digital Imaging
Define Significant Properties
Analyze embedded metadata, what metadata should augment what exists in DRS?
Analyze content model, is metadata sufficient?
Research available toolset, score acceptability
Consider toolset's ability to extract metadata, analyze characterization tools
Roles
Responsible
Technical Support
Accountable
Consulted
Informed

Figure 2: Excerpt of Migration Workflow Diagram with example of RACI color-coding for Stakeholder involvement

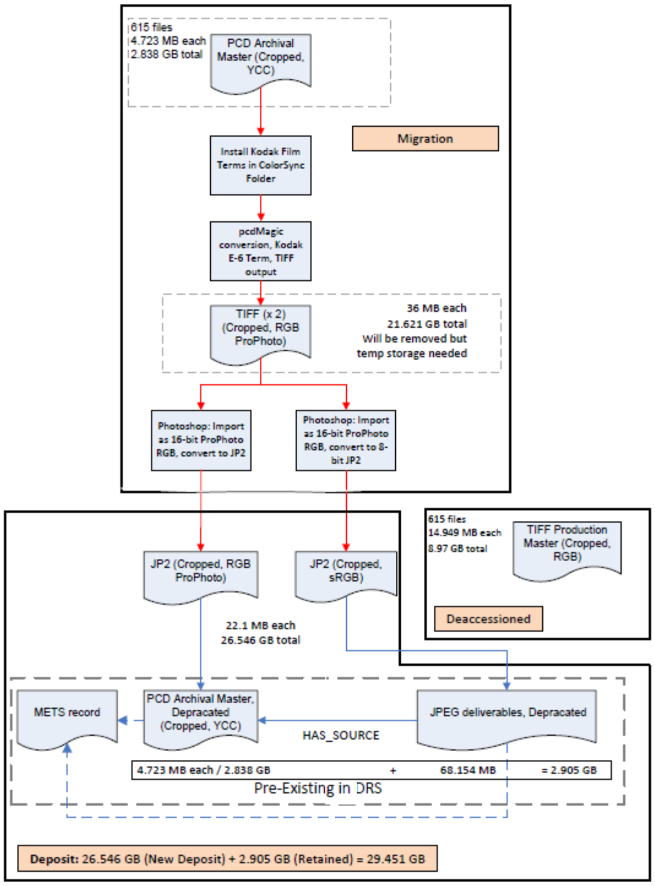

615 files
4.723 MB each
2.838 GB total
PCD Archival Master (Cropped, YCC)
Migration
Install Kodak Film Terms in ColorSync Folder
pcdMagic conversion, Kodak E-6 Term, TIFF output
TIFF (x 2) (Cropped, RGB ProPhoto)
36 MB each
21.621 GB total
Will be removed but temp storage needed
Photoshop: Import as 16-bit ProPhoto RGB, convert to JP2
Photoshop: Import as 16-bit ProPhoto RGB, convert to 8-bit JP2
JP2 (Cropped, RGB ProPhoto)
JP2 (Cropped, sRGB)
22.1 MB each
26.546 GB total
615 files
14.949 MB each
8.97 GB total
TIFF Production Master (Cropped, RGB)
Deaccessioned
METS record
PCD Archival Master, Depracated (Cropped, YCC)
HAS_SOURCE
JPEG deliverables, Depracated
4.723 MB each / 2.838 GB
+
68.154 MB
= 2.905 GB
Pre-Existing in DRS
Deposit: 26.546 GB (New Deposit) + 2.905 GB (Retained) = 29.451 GB

Figure 4: Excerpt of Migration Pathway Diagram for the Harvard Daguerreotype/Richard Ree Collection. The top box shows the migration process that will start with a PCD Archival Master, produce an intermediary TIFF image, and then will be converted to two JP2 images. The bottom box shows that the two JP2 images will be deposited to the DRS and related ot existing DRS images. Migration processes are shown as red lines; documented DRS metadata relationships are shown as blue lines.

A most essential final step in this phase is to finalize the definitions of the migration environment – the computational systems, storage processes (both temporary and permanent), key hand-offs of content throughout the workflow and tool set-up and use requirements. This list will help to expose any additional development that may need to take place on the existing technology. These needs should be considered in the overall Migration Plan and Workflow Diagram though are detailed specifically in **Batch Ingest** and **Metadata Mock-Ups**.

For the PCD plan, it was decided that in all cases two JPEG2000 JP2s would be created and would replace previous master and derivative files, one with the RGB ProPhoto color space to serve as an Archival Master (PCD as back-up), and one with the sRGB color space to serve the purposes of both Production Master and Deliverable. It is worth noting that while a TIFF is generated from the pcdMagic tool that the JP2 is the ultimate target format that will be saved from the migration. In the case of the Harvard Daguerreotypes and Richard Ree projects, the cropped PCD Archival Master would serve as the source for both the ProPhoto RGB JP2 and the sRGB JP2. For the Horblit Collection the uncropped Archival Master would serve as the source for the ProPhoto RGB JP2 and the cropped Production Master would serve as the source for the sRGB JP2.

Upon ingest of the JP2 files, new relationships will need to be added to link the JP2 images to the source files that they are replacing. The TIFF Production Masters from the Harvard Daguerreotypes and Richard Ree Collections may not be retained since as described earlier they were generated using inferior software that did not account for the SBA settings. Though new JP2 deliverables are being created as part of the migration, the older JPEG deliverables need to be retained since they have persistent names (URNs) published in catalogs and web pages. The original PCD images will be kept in the unlikely event that a future migration effort is performed (with newer, better tools on the market, which is also highly unlikely). The PCD images are relatively small so they do not affect storage capacity too greatly.

3.4.4 Phase 4: Execute Plan

3.4.4.1 Sub-Phase 1: Schedule Migration

In this sub-phase the migration project is scheduled and staff resources for the migration execution are assigned.

At the time of this writing the PCD migration has not been scheduled yet. This project was being done as an NDSR residency project, and the residency term ended after nine months, putting the project on hold. The project remains a high priority but will have to wait until there are staff resources within Digital Preservation Services that can continue this work as this department is taking the lead on the project.

3.4.4.2 Sub-Phase 2: Custom Development

Especially for the first migrations within an organization, they will likely require custom development by the Systems and Technology Experts. In some cases new scripts will need to be created to create a migration pipeline in which conversion tools can be inserted and removed as needed, in other cases existing tools will need to be modified.

In the PCD case Library Technology Services will need to modify its DRS ingest tools to be able to add the files created through the migration to the existing DRS objects. The existing DRS deposit tools can only add new objects to the DRS, not modify existing objects. This is indeed an issue since some products of the migration will replace content previously contained within the image object (e.g. TIFF intermediate files that were created using inferior conversion tools/processes which will be replaced by new JP2 files). In addition they will create a script so that pcdMagic can be called programmatically.

3.4.4.3 Sub-Phase 3: Conduct Migration

This is the sub-phase where the actual migration is conducted. It concludes exporting the content that will be used as the source of the migration to a temporary storage area, conducting the migration according to a **Migration Checklist**, and depositing the content to the DRS.

In the case of the PCD migration, the source PCD images and their associated METS metadata files will be exported by Library Technology Services to a directory structure specified by the Analyst. The values of specific metadata elements (methodology, role, and relationships) will be used by the migration tools to know which parameters to use and which files to create.

3.4.5 Phase 5: Verify Results and Wrap-Up Project

After the migration and ingest to the DRS there will be need for the final checks, documentation and clean-up. The metadata and reports that are generated throughout the workflow should be re-checked to confirm the success of key processes, that the migration was complete and that the metadata and content results are as expected and documented in the **QA Report**.

This is also the sub-phase where the de-accessioning plan developed earlier in the workflow is revisited to see if additional steps need to be taken, for example if files should be deleted or simply made inactive. This is also the appropriate stage for reviewing all the documentation produced throughout the migration. Ensuring that each document accurately reflects the final process is very important as these will likely be referenced for future migration projects as well as serving as authoritative provenance documents for demonstrating the chain of custody of the content. At this point it should be decided if any of these documents merit inclusion in the repository along with the files. The framework ends with writing any lessons learned in a **Migration Conclusions** document to inform future migrations.

4. CONCLUSIONS

While the Kodak PhotoCD and RealAudio/SMIL Playlists migration plans are still underway, simultaneous development of the plan for each format and the generic migration framework has helped to conceptualize the process for each format, identify aspects common across the format plans, and provide more certainty that a generic framework is possible. While the framework is very specific to the processes and procedures at Harvard Library, it is hoped that the framework will be helpful to other institutions as they approach migrations as a preservation action for their digital collections at scale.

5. ACKNOWLEDGMENTS

We would like to thank David Pearson of the National Library of Australia and Peter McKinney of the National Library of New Zealand, both of whom generously shared drafts and finalized versions of documentation and workflows from their format migration projects. We would also like to thank the Institute of Museum and Library Services for making this project possible through the funding of the NDSR Boston residency project. After finishing his residency at Harvard, Joey Heinen accepted the Digital Production Coordinator position within the Scholarly Communication department at Northeastern University.

6. REFERENCES

[1] Hutar, J. 2013. Assessing Digital Preservation Strategies. Archives New Zealand (Wellington, New Zealand). http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00155.pdf

[2] National Digital Stewardship Residency (NDSR-Boston). 2015. http://projects.iq.harvard.edu/ndsr_boston

[3] Harvard Library. “Overview: DRS & Delivery Systems”. http://hul.harvard.edu/ois/systems/drs/

[4] PREMIS Editorial Committee. 2011. Introduction and Supporting Materials from PREMIS Data Dictionary for Preservation Metadata, version 2.1. http://www.loc.gov/standards/premis/v2/premis-report-2-1.pdf

[5] Gattuso, J. and McKinney, P. 2014. Converting WordStar to HTML4. In *iPres 2014* proceedings. Archives New Zealand (Wellington, New Zealand). http://ndha-wiki.natlib.govt.nz/assets/NDHA/Publications/2014/WordStar-ipres2014-4.pdf

[6] Schaller, M. and Schlarb, S. 2014. SCAPE: Large Scale *Research and Development Department* (Vienna, Austria). https://onbresearch.wordpress.com/2014/06/16/scape-large-scale-image-migration/

[7] Bajcsy, P., Kooper, R., Marini, L., McHenry, K. and Ondrejcek, M. 2010. A Framework for Understanding File Format Conversions. In proceedings for *Roadmap for Digital Preservation Interoperability Framework Workshop*. University of Illinois at Urbana-Champaign (UIUC) (Champaign, IL). http://dl.acm.org/citation.cfm?doid=2039274.2039284

[8] Becker, C., Kulovits, H. and Rauber, A. 2010. Trustworthy Preservation Planning with Plato. In *European Research Consortium for Informatics and Mathematics, Is. 80*. pp. 24-25. Technical University Vienna/AARIT (Vienna, Austria). http://ercim-news.ercim.eu/images/stories/EN80/EN80-web.pdf

[9] Luan, F., Nygard, M., Sindre, G., and Aalberg, T. 2011. Using a Multi-Criteria Decision Making Approach to Evaluate Format Migration Solutions. In conference proceedings for *MEDES ’11*. San Francisco, California. http://dl.acm.org/citation.cfm?doid=2077489.2077498

[10] Burns, P., Madden, T., Girogianni, E. and Williams, D. 2005. Migration of Photo CD Image Files. In conference proceedings for *IS&T: The Society for Imaging Science and Technology*. East Kodak Company (Rochester, New York). http://losburns.com/imaging/pbpubs/43Arch05Burns.pdf

[11] Jack, K. (2005). Color Spaces. In *Video Demystified: A Handbook for the Digital Engineer* (4th Edition, pp. 15-34). Elsevier. http://www.compression.ru/download/articles/color_space/ch03.pdf

[12] McGuffog, S. 2015. pcdMagic User Manual. https://sites.google.com/site/pcdmagicsite/

[13] Felix, T. 2009. Software the *Really* Supports Kodak Photo CD. http://tedfelix.com/PhotoCD/PCDSoftware.html

[14] Eastman Kodak Company. 1992. Image Pac Compression and JPEG compression: What’s the Difference?

http://www.kodak.com/digitalImaging/samples/imagepacVsJPEG.shtml

[15] Kodak Professional. 2000. Using the ProPhoto RGB Profile in Adobe Photoshop v5.0. http://scarse.sourceforge.net/docs/kodak/ProPhoto-PS.pdf

[16] Hill, B., Roger, T. and Vorhagen, F.W. 1997. “Comparative analysis of the quantization of color spaces on the basis of the CIELAB color-difference-formula. In *ACM Transactions on Graphics, Vol. 16 Is. 2* pp. 109-154. http://portal.acm.org/citation.cfm?doid=248210.248212