9-4-2021

Adventures in Migrating Massive Archival Collections to Digital Commons

Michelle Emanuel
University of Mississippi, memanuel@olemiss.edu

Recommended Citation

Emanuel, M. (2021). Adventures in Migrating Massive Archival Collections to Digital Commons. Open Information Science, 5(1), 119-123. https://doi.org/10.1515/opis-2021-0007

Communication

Michelle Emanuel*

Adventures in Migrating Massive Archival Collections to Digital Commons

https://doi.org/10.1515/opis-2021-0007
received June 30, 2020; accepted June 13, 2021.

Abstract: Giant collections of compound objects, with messy metadata, led to migration in batches and lessons learned. When the University of Mississippi Libraries implemented a campus-wide institutional repository, it also became necessary to use the same platform for nearly 100 digital collections, requiring migration from a locally hosted instance of CONTENTdm to the cloud-based Digital Commons. Because the collections were primarily comprised of compound objects, it was difficult to use harvesting protocol to populate the new repository, requiring new access copies to be reconfigured and uploaded as batches from the local servers after extensive metadata remediation.

Keywords: Institutional Repositories; migration; digital collections; CONTENTdm; academic libraries

1 Introduction

In 2018, the University of Mississippi Libraries (UML) convened a task force to find a suitable platform to launch a hosted institutional repository (IR). Prior to the task force's exploratory stage, OCLC announced that any local instances of CONTENTdm (CDM) would need to replaced before the end of 2018; for UML, this meant we would also need to migrate our digital collections to either a hosted version of CDM, or to a new platform. In an effort to minimize costs, we were challenged to find one platform to serve both functions: launch an IR and both maintain and increase our digital collections. We selected Digital Commons from bepress and named our instance eGrove, in honor of the 10-acre green space at the center of our campus. The bigger adventure came after the IR's launch: the migration of our legacy content, totaling over 48,000 items in 88 collections, from CDM to Digital Commons (DC). We faced a number of obstacles in the process, some expected and others unexpected. For one, CDM uses "compound objects" to represent multi-paged items. Depending on the collection's configuration, the initial PDF of a document can be split into hundreds of single-page files. Though we knew that DC does not use compound objects, we had not anticipated how difficult it could be to locate the original access copy for the new ingest. Also, after the metadata was initially remediated to control shared vocabularies, many had to be adjusted again to conform to DC templates, particularly fields related to names and dates. We also had to work with bepress Consulting Services to customize our metadata fields to allow for non-standard fields that had been present in our CDM records, such as Library of Congress Subject Headings (LCSH) and Art and Architecture Thesaurus (AAT) terms.

Though we had ample notice from OCLC that our local server would no longer be supported after the end of 2018, it did not make the process of migration any easier. From 2006 to 2016, our digital initiatives program had been a decentralized effort with content generated by the individual curators in Archives and Special Collections (ASC) and our Accountancy librarian. The Digital Initiatives (DI) librarian, at the time part of the ASC unit, managed the CDM administrative module in cooperation with UML's Department of Library Technology (Library IT), with additional input from the Web Services (WS) librarian. The DI librarian would receive the digital objects with accompanying metadata from the curators and upload it into the appropriate CDM collection. Library IT would manage the servers housing the access and master copies of the digital objects plus backups. The WS librarian managed the display of the digital

***Corresponding author: Michelle Emanuel,** University of Mississippi, Oxford, Mississippi, United States, E-mail: memanuel@olemiss.edu

collection within the local instance of CDM. The Accountancy Librarian, digitizing an extensive number of documents of the American Institute of Certified Public Accounting (AICPA), was adding items one at a time to four established collections using the template-based local client. As long as he did not deviate from the workflow, the DI and WS librarians did not need to intervene. Initially, ASC used the local client to upload its smaller collections, as the template worked well with copying and pasting, and was easy to delegate to student labor. As the collections grew larger, however, batch uploading with spreadsheets through the administrative module became the preferred ASC method, primarily to maintain consistency and controlled vocabularies. The Accountancy Librarian, however, continued to use the local client. In 2016, the Cataloging unit, which had only served in an advisory role, started to look over the spreadsheets before ingest, leading to a large retrospective metadata cleanup project (Ivey, Emanuel, 2018). But because there was no easy way that we could determine to batch revise collections in CDM, the decision was made to save the revisions for an eventual migration to a new platform.

The metadata cleanup project took approximately two years, mostly focusing on bringing the collections into compliance with the metadata best practices document adopted in 2016. Using Open Refine to isolate inconsistencies in the datapoints, we looked at each of the metadata fields in each of the CDM collections. Recurring issues included errant punctuation in the Title field, name variants in the Creator and Contributor fields, and incomplete Date fields. In DC, Creator and Contributors become "author 1", "author 2," etc. and would each need reformatting from [lastname, firstname] in one cell, separated by semicolons, to separate columns for first name, middle name or initial, last name, and suffix. The date would also need to be reformatted to yyyy-mm-dd. CDM fields for Original Format, Access Format, and Master Format (dc: format) adopted the Getty Institute's Art and Architecture Thesaurus as its controlled vocabulary while the Type field was simplified to six choices: audio, image, interactive resource, physical object, text and video. The Original Collection Name (dc: source) and Digital Collection Name (dc: RelationIsPartOf) were standardized. Measurements in the Extent field (dc: format-extent) were simplified by replacing apostrophes and quotation marks with "ft." and "in.", and limiting the number of decimal places to two. Finally, the URLs for the collection finding aids were updated. By looking at each field of each record in each collection, we were able to identify which collections would be easy to migrate – because there were no compound objects, and no format variations – and which would require reorganization.

Unfortunately, as we were deep cleaning the metadata, we did not spend any time looking into the location or format of the collections' access copies, which were scattered among several shared drives managed by Library IT. In retrospect, this is where we should have started the cleanup project. We quickly realized that our new DC structures would need to be set up according to format. A collection in CDM could include a variety of formats and file types, but a DC structure had to be either a book gallery or image gallery if a thumbnail image should display. To decide which collections to migrate first, we prioritized the list based on CDM site analytics. Unfortunately, this process revealed that our most massive collections were our most popular: the AICPA collections, which had received a lot of referrals from an earlier linking project in Wikipedia, our collections for Civil Rights, the Civil War, and general photographs. Though we did upload some smaller collections to help learn the system gently, we could not avoid diving into the deep end. The AICPA collections combined contained over 15,000 items, mostly PDF files on one shared drive, though unfortunately without a consistent naming convention. Knowing which file was the "final version" was made more complicated by AbbyyFineReader, a program used to check each file's OCR, which had generated multiple folders and versions. In CDM, the PDFs appeared as "compound objects"; the uploaded PDF would split into individual pages that would had to be clicked through one by one, using the side bar. On the other hand, the ASC collections, stored across five shared drives, included photographs, manuscripts, as well as sound and video recordings. Both photographs and manuscripts had been scanned as TIFF images, which CDM uploaded as JPEG2000 files, also called .jp2 files. But while many manuscripts rendered as jp2 files, others were read-only pdfpage files, with each page of the manuscript having its own file. Like the AICPA collection, all of the ASC manuscript collections had compound objects. Unlike the AICPA collection, where each item had a master PDF to migrate, the manuscript collections had two scenarios: 1) single compound objects, and 2) thousands of single pages. Due to the thousands of compound objects with varying file types across our collections, typical harvesting by bepress was going to be incredibly difficult, and prone to error. We made the decision to handle the migration ourselves, using the Batch Upload function for each collection.

We soon discovered that DC would not display .jp2 files, and the thousands requiring migration have to be converted to JPG or PNG; the choice of display structure in DC would determine whether those images stayed as image files, in the case of a photograph collection, or be combined to make PDFs for a manuscript collection. We also had to

figure out where the files – both access copies and master copies – were living, in order to manipulate them. Though understandable from a security standpoint, the extensive set up of access and permissions for the different shared drives hindered project delegation.

Less than six months before our migration started in earnest, in a presentation at the 2018 Digital Initiatives Symposium in San Diego, Kristin Laughtin-Dunker talked about Chapman University's migration from CDM to DC using what she termed a "FrankenURL". By manipulating the original object's URL – specifically, changing "cdm" to "utils" and "compoundobject" to "getfile" – a new URL could be generated for the batch upload. (Gibney 2018)

> Original URL: http://clio.lib.olemiss.edu/**cdm**/**compoundobject**/collection/deloitte/id/3714/rec/1
> FrankenURL: http://clio.lib.olemiss.edu/**utils**/**getfile**/collection/deloitte/id/3714
> (*emphasis mine*)

We could not wait to try this method, but unfortunately in our experience, it proved inconsistent. For whatever reason, in any given batch, less than 100% would successfully upload. We would have to see where on the spreadsheet the upload stopped, then copy the missing items to a new spreadsheet, and try again. In a book gallery structure, sometimes PDFs would upload without grabbing the cover image. In those cases, the cover image would then have to be saved as a JPG or PNG, and uploaded by hand or as part of a batch edit. In image gallery structures, many of our files were rather large and would time out before finishing the upload process. As with the PDFs, we would have to look at the original batch spreadsheet to figure out where the process had stopped. And because we were uploading hundreds of files to multiple structures, trying to accomplish as much as possible before our CDM instance on a temperamental local server fell into the dreaded "unsupported zone", we had to accept rather quickly that this method was not going to work for us.

After much experimentation, Library IT designated space on a shared server where we could place folders of access copies for any particular batch upload. We would then notify them that a new folder had been added. They then posted the folder(s) to the same server that hosts the library's website just long enough for the files to be uploaded to DC. From the web server, Library IT could export a list of URLs, which we could put into the "fulltext_url" field of batch upload spreadsheet. We soon realized that we could also use this method for the "cover_image_url" field. Any joy we might have felt from this innovation, however, was short lived. Not only would compound objects be a major roadblock to our progress, but there were actually two types of compound object to contend with. The first type is the "simple compound"; in the CDM metadata export, there is a separate line for each page of the PDF, saved as a jp2 file, with another line at the bottom of the stack for the compound object itself, a cpd file, containing the descriptive metadata.

	A	B	C
1	CONTENTdm	Identifier	Title
2	1.jp2		clark_b1f2_001
3	2.jp2		clark_b1f2_002
4	3.cpd	clark_b1f2	T.G. Clark to Margery Clark (9 December 1861)
5	4.jp2		clark_b1f3_001
6	5.jp2		clark_b1f3_002
7	6.cpd	clark_b1f3	T.G. Clark to Margery Clark (13 December 1861)
8	7.jp2		clark_b1f4_001
9	8.jp2		clark_b1f4_002
10	9.jp2		clark_b1f4_003
11	10.cpd	clark_b1f4	T.G. Clark to Margery Clark (23 December 1861)
12	11.jp2		clark_b1f5_001
13	12.jp2		clark_b1f5_002
14	13.jp2		clark_b1f5_003
15	14.jp2		clark_b1f5_004
16	15.cpd	clark_b1f5	T.G. Clark to Margery Clark (28 December 1861)

Figure 1: Simple Compound Objects.

The jp2s can be converted to JPG or PNG, then combined to make a new access copy, saved as a PDF with the same identifier as the master files. This was definitely time consuming, especially for items with lots of pages on long spreadsheets, but was relatively easy for the eye to follow. Much more challenging were the second kind of compound object: the pdfpages. In these assets, each page of the original PDF upload saved as a separate file, in a format that could not be manipulated in Adobe Acrobat.

	A	B	C	D	E	F
1	CONTENT	Identifier	Title			
2	11436.jp2	mum00400_b1996-1_f07_016	Elizabeth Christie Brown Diary: Scan 16			
3	11437.jp2	mum00400_b1996-1_f07_017	Elizabeth Christie Brown Diary: Scan 17			
4	11438.jp2	mum00400_b1996-1_f07_018	Elizabeth Christie Brown Diary: Scan 18			
5	11439.jp2	mum00400_b1996-1_f07_019	Elizabeth Christie Brown Diary: Scan 19			
6	11440.jp2	mum00400_b1996-1_f07_020	Elizabeth Christie Brown Diary: Scan 20			
7	11441.jp2	mum00400_b1996-1_f07_021	Elizabeth Christie Brown Diary: Scan 21			
8	11442.jp2	mum00400_b1996-1_f07_015	Elizabeth Christie Brown Diary: Scan 15			
9	11443.jp2	mum00400_b1996-1_f07_023	Elizabeth Christie Brown Diary: Scan 23			
10	11444.jp2	mum00400_b1996-1_f07_024	Elizabeth Christie Brown Diary: Scan 24			
11	11445.jp2	mum00400_b1996-1_f07_025	Elizabeth Christie Brown Diary: Scan 25			
12	11446.jp2	mum00400_b1996-1_f07_026	Elizabeth Christie Brown Diary: Scan 26			
13	11447.jp2	mum00400_b1996-1_f07_022	Elizabeth Christie Brown Diary: Scan 22			

Figure 2: pdfpages.

Each page had its own pdfpage, displayed separately in CDM; there was no cpd file. And, to make matters worse, the files did not display in order. Looking at the CDM file names in ascending order, there was at least one outlier in any given document to throw off the sequence. In this situation, it was ultimately easier to go back to the master TIFF images in dark storage, where the identifiers could be lined up in order, converted from TIFF to PNG, then combined in Adobe Acrobat to make a new access copy.

Finally, there were thousands of photographs to migrate. These were not compound objects, and could be easily uploaded using the method we had arranged with Library IT except for one detail: the filename of the watermarked access copy, which CDM had rendered as an accession number, no longer matched the identifier in the metadata. This tiny detail was going to prove difficult for the Audio & Video Collections Librarian anytime a patron might ask about licensing an image for publication, if the numerical file name were provided instead of the identifier. Fortunately, we found a solution in a presentation from Amanda Mita and Zachary Peli at the 2018 Mid Atlantic Digital Commons User Group about their own migration from CDM to DC, in which they had manipulated a Python script to retrieve images, remove duplicates, and rename files. Fortunately, our in-house Collection Applications Developer was able to similarly manipulate Python to extract the watermarked access copies from CDM, rename them to match the identifier, and group them in a folder on a shared drive to give to Library IT for uploading.

With the access copies in hand and the metadata cleaned up, we were finally ready to batch upload each collection. All told, the migration took just under a calendar year, meaning that we were relying on an unsupported platform for most of the year. Though this risk made us extremely nervous, it did not prove to be as problematic as we had feared. In the end, it was enlightening to determine the exact size of our digital collections: how many files did we actually have online? How many files in our dark archives were actually duplicates? Or were scanned but never posted online? When calculating our preservation needs, both in server space and cloud storage, our overall file count had been based on an inflated page count rather than a number of actual documents. Future projects include auditing the size of the files in our dark archive, deleting unnecessary duplicates, and finding long term storage solutions for our digital collections.

Funding Information: Authors state no funding involved.

Conflict of Interest: Author states no conflict of interest.

Data Availability Statement: Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

References

Gibney, M., Laughtin-Dunker, K., and Chance, E. (2018, April 24) Migratory patterns in IRs: CONTENTdm, Digital Commons, and flying the coop. Paper presented at Digital Initiatives Symposium (University of San Diego, San Diego, California, USA). https://digital.sandiego.edu/symposium/2018/2018/34/

Ivey, S. and Emanuel, M. (2018) "Large scale with a small staff and even smaller budget: Updating metadata to reflect revised best practices." *Organization, Representation, and Description Through the Digital Age: Information in Libraries, Archives and Museums*. Eds. Caroline Fuchs and Christine Angel. DeGruyter, 241-254. DOI: 10.1515/9783110337419-017

Mita, A. and Pelli, Z. (2018, July 27). Migrating from CONTENTdm to Digital Commons: Considerations and Workflows. Paper presented at Mid Atlantic Digital Commons User Group (Benjamin N. Cardozo School of Law, New York, New York, USA) https://larc.cardozo.yu.edu/madcug/2018/program/12/