Electronic Thesis and Dissertations in Acquisitions

UMBC has some ContentDM Digital Collections and includes with them types of materials that fit better in the Intuitional Repository Category. However, UMBC doesn’t consider ContentDM an Institutional Repository software, and on investigation, we found it unsuited to serve as one. University of Utah does use Contentdm as a repository, but only with programming added on to it to make it work as one. OCLC is currently working with user groups to identify features lacking that prevent it from being used as one, but has not made a commitment to adding features for that purpose.

A primary emphasis of UMBC’s Special Collections is photography and the history of photography, and we had digital images only before we purchased Contentdm in 2006. Contentdm was originally purchased for digitized photos and we didn’t intend to use it for other purposes but have done so since it’s the platform we currently have available. I’m going to show you some images from our original collections which went live in the fall of 2009 so you’ll understand what our main purpose and use for ContentDM.
Civil War Collection

This is an image from our Civil War Collection of Harper’s Ferry during the Civil War.
Bretz Collection

This is our Bretz Collection, photographs of coal mines and coal miners.
Hines Collection

This is an image from our Hines Collection, which are photographs of child labor. There are children working for a packing company. All of our photos were digitized in Special Collection. Many were digitized and available in html before we purchased contentdm. Staff responsible for bringing up our digital collections ContentDM are Susan Graham, Special Collections Librarian, and Lindsey Loeper, University Archivist, And in Technical Services, cataloger Vicki Sipe who helped with the metadata. Web Designer Janet Hack designed our public Contentdm website which went live in 2009.
University Archives

We also have digital University Archives, part of which is digitized print collections, but some is being sent to us in a digital format for inclusion in the Digital Archives.
GEST (Goddard Earth Science Technology Center) publications are examples of materials being sent to us in a Digital Format for inclusion in our Digital Collections. These types of materials straddle our definitions of Digital Collections and Institutional Repository, in that they are already in our collection in print, or would be included in print, but they are materials being submitted to us electronically by the author for us to provide access and preservation.
We also have Electronic Thesis and Dissertations in our Contentdm Digital Collections, although most usually ETD’s are considered part of an IR. I’ve talked to a library that currently utilizes both Contentdm and an IR software, and they’ve indicated that there are materials that are particular suited to one or the other of the systems, but in instances where the line is blurred between the two, they chose based on what is easiest and works best.
As we began to bring our Digital Collections up in February 2009, I proposed that my Acquisitions Unit take responsibility for Transfer Related Service, which I’d read about in a blog post. I later reconceived this as Digital Transfer Services. I began talking with Special Collection regarding materials coming into ContentDM and support Acquisitions might be able to provide. My concept of what Digital Transfer services is includes:

- Acquiring digital content
- Conducting quality review of digital content
- Moving digital content between systems
- Inventorying
- Manipulating
- And ingesting it into the appropriate application or repository.
I had in mind that we could work with the e-serial files being periodically acquired by receiving the files, inventorying them, checking them, and then loading them into ContentDM.
But the problematic workload at the time was ETD's, which were received from Proquest along with metadata for them, and this was what they wanted to give me. The librarian who had been doing them had resigned, the programming that reformatted the metadata had broken and we no longer had a programmer to fix it. Special Collections took the work temporarily and was having a student assistant manually reformat the data. The work was supposed to shift back to the position it had come from when that position was filled but the workload for that position was already heavy so the work came to Acquisitions instead.
What I knew about workflow when I agree to take it over was that we had ETD’s in ContentDM with metadata that looked like this,
And ETD’s in PDF format that looked like this. I went in and looked at the collection with a mind for understanding what was going on with them, how they would work, and how patrons access and view them. I found some issues and Lindsey Loeper, our archivist doing the work, made me aware of some other issues.
Initial Tasks

- Automate the movement of XML metadata into the format needed for contentdm, tab-delimited text files.
- See if the grad school could do anything to stop the students from including personal info in the CV's included with the ETD's.
- See if the grad school could provide a list of everything we should be receiving from Proquest.
- Find a way to get links into the records in the catalog.
- Correct ETD values to match those of other collections where appropriate.
So, my first step was to talk with a bunch of people and get information.

- I found from our ContentDM experts that the data inconsistency in ContentDM could and should be resolved with find-replace function in ContentDM. That remains on my list of things to do.

- I found from my Head of TS, Lynda Aldana, about the lack of links in the catalog records, and was told that after I completed this part of the project she'd like me to look at loading the ContentDM records to OCLC and to Aleph so that links would be included and that we would no longer have to catalog them.

- I found from the grad school that they were also concerned about personal info in the CV's being published with the ETD's and they agreed to modify their template.

- The Grad School also agreed to inform me of materials we should expect to receive (but haven't as of yet done so).

Additionally, they requested that the library begin paying for print copies of the CV's for special collections. That request went up the chain of command and was agreed to, and so we began also receiving and paying for the print ETD's. It turned out the grad school had a large backlog of print ETD's as well as unopened invoices from Proquest which we went and fetched from them and sorted through and resolved. Future shipments and invoices will come to the library directly and will be received and invoiced in Acquisitions.
So it appeared that two of the items on my list were completed, and I added a new task of take over the receipt and payment of hardcopy thesis and dissertations. However, neither of the things agreed to with the grad school have occurred yet. I’m hopeful changes will occur soon, but it may take some follow-up.
Next I needed to add 3 new pieces of software to my computer, and I added a 4th later when I began looking at the existing procedures. The software needed was FileZilla, an FTP tool for downloading the files, 7-Zip File Manager, a utility for unzipping and moving zip files, Adobe Acrobat, a tool for manipulating .pdf files, and the ContentDM client, used for working with materials to load onto the ContentDM server. The ContentDM also needed to be configured, which was initially done by Lindsey, our Archivist, and Lindsey also provided me with a written procedure outlining how the ETD’s were being processed at the time, and walked me through the procedure once.
The process works like this. We receive e-mailed notices from Proquest that look like this that we have new files. If you look up at the top of this notice, you’ll note that indicates that 29 of 53 files delivered successfully, and 24 of 53 failed to deliver. Usually, the files successfully load, but other times they don’t, and thereafter they sometimes come in small bits. This can make it challenging to figure out if you received all of the files or not except by detailed checking for each of the failed loads item-by-item in later notifications.
When all the files are on the server, we download using Filezilla, an FTP utility. This is a simple drag and drop utility.
The downloaded files are now on my computer in zip folders. The procedure had been to open each individually and drag and drop each pdf and xml file to the appropriate folder to separate them, but this is what I got the unzip program for.
7-Zip and Supplemental Material

In 7-Zip I can find the directory the zip directories are in, highlight them all, then click extract. When you go back to look at the directory and refresh, the pdf and xml files are all in the main directory, so the files can be sorted by type then dragged and dropped as a group. As we look through and sort the files received, most have .xml file and a .pdf file, but others come with the authors name on them and include extra files that came with the .xml and .pdf files. These may be approval sheets or viteas, which are usually included in the main .pdf, in separate files. In other cases, they may be supplemental materials, sometimes that aren’t included in the main .pdf already, but sometimes they are included already. Sometimes supplemental material files will be in MACOS. If supplemental materials aren’t included, files need to be combined into the main .pdf using adobe, or alternatively, if this won’t work because of the nature or format, they need to be attached to the main object in ContentDM.
Excel Appendixes Spreadsheet

In this instance this dissertation included datasets in a number of Excel files with genetic codes and I couldn’t find a way to convert this into a usable .pdf.
Excel Appendixes Spreadsheets in Contentdm

So I ended up attaching in Excel format for user download. In an instance like this where this isn’t a clear and documented method for handling, the staff doing this work sends all of the files to me to handle. In simpler instances where materials that are usually together in the main .pdf but are separated, or there is a supplemental .pdf that can be combined with the main .pdf. procedures are in place for staff to handle.
Missing ETD’s
Once we have the extra files figured out, we have to open all of the ETD’s to delete the CV’s because of the personal information in them. In some instances when we’ve opened the PDF file to delete the CV, we’ve found that there is no thesis or dissertation included in the file, but only the approval sheet. Initially I searched these in Dissertation Abstracts to confirm that they all had embargoes on them, but there were also missing files that weren’t embargoed. At this point I contact Proquest for more information and found that I could determine if there were embargoes on items from the metadata, as well as the date when the embargoes ends. Additionally I found out that the missing files without embargoes were too large for the authors to upload to Proquest so had been sent on disk via mail, in which case we don’t get the files automatically but have to go to a Proquest Website to retrieve them.
Embargos

The embargoed document have an abstract which we load the embargoed documents into ContentDM without the document, and expect to receive and load the full document later. We maintain a list of embargoed ETD’s to check on them later. When I checked on this one I found that we had received and loaded the full ETD, so that the partial one now needed to be deleted.
So at this point, we’ve Received notice from Proquest that there are files available, FTP’ed files, Unzipped files, Sorted files into folders for XML and for PDF’s, Figured out what to do with extra files, Noted embargoed documents on our “embargo list,” and we’ve retrieved missing files from the UMBC Proquest website.
Next we need to get the metadata we need from the XML files into an Excel spreadsheet that can be saved as a text-delimited file for import into ContentDM. This was being done by manually locating each needed piece of data in the individual XML file then copying and pasting it into the spreadsheet. I imagined this to be very time-consuming and tedious for just 1 file much less for 50 or 60,
Metadata Map

So I utilized the materialize Lindsey had given me to make an easy to read map for the metadata and it’s relationship to the XML and began to automate using Excel advanced operations and Excel XML Tools and Macro Developer Utilities. While the process didn’t end up quite as effortless as I wanted, it came up vastly better way than copying and pasting bit by bit. The metadata conversion process works like this:
Excel will automatically map metadata for you when import the first XML file of a particular type. It maps all of the metadata to the spreadsheet, and in the order of the XML. We don’t need all of it, and needed it in the ContentDM order, so I modified Excel’s mapping. This is done by clicking source. In the XML source map, you click any bolded element to see where it’s mapped to. The yellow highlighted field is where the author surname maps to. By right clicking an element you can remove it, and you can drag and drop any element to any cell on the spreadsheet and it will mapped to that spot. All of the brown fields in the template spreadsheet or directly mapped from the XML.
In many instances data is separated in the XML that needs to be combined for contentDM. An important function for doing this is the Concatenate function. This can be done either using a function that begins =concatenate as shown above, or with an & to join text terms. Characters you want to appear between the text terms are put in quotes. For example, the above function puts together XML mapped metadata for author last name, first name, and middle initial, inserting a comma between the last and first names and a blank space between the first name and middle initial.
I used macros to change values and manipulate data in many different ways. Most commonly I used find and replace macros or copy and paste macros as shown above. Macros can be recorded and written in visual basic. I generally use a combination of the two. I'll begin by recording, then edit to add on to that macro. For things like find/replaces, it's very simple to copy a find/replace macro and change the values to run an additional find and replace. Other macros may be more difficult to edit such as the copy/paste at the bottom with a pastespecial with variables set appropriately. Sometimes, instead of completely recording or editing, I'll re-record the snippet that needs correction, then copy the visual basic for that snippet and paste it over the part of the macro that needs to be fixed.
Excel Deletions

Open the ETD Metadata Template Excel file. Go the ContentDM tab and push CTRL-X to remove all old content...
Proquest Tab

Go to the Proquest tab and
S-30--XML Import

Click Developer and import and select the first XML file to import
Concatenate

At the bottom of the spreadsheet, cells with concatenate functions and macros combine multiple advisor and dissertation paragraph fields without the user doing anything.
At the top of the spreadsheet, a variety of simpler Excel functions reformat data in a variety of ways, again without the staff person doing anything. The staff person just runs a “mover” macro by pushing CTRL-M
Into Content DM Worksheet

That moves the data into the ContentDM sheet
Move and Move and Move

While the user remains in the Proquest sheet continuing to Import, then CTRL-M each file. When all the files are loaded and moved, the user then goes to the complete ContentDM load file.
S-35--Contentdm Sheet Ctrl-R Reformat

And pushes CTRL-R to further reformat the data to be compatible with our ContentDM data.
Formatted Data except Keywords

All data ends up in the proper form without additional staff intervention with the exception of the keywords
Keywords

Which come from ProQuest separated by commas. These are user input, so sometimes the keywords include commas in them. For ContentDM, we need the keywords separated by semi-colons, but for the commas within a single keyword to remain commas. So we find an replace to do this, with staff checking each replace before doing it to insure that it’s appropriate. If keywords include a value like
Keyword Examples

For example, if the keywords contain this, the staff member clearly leave the commas between Frederick and Maryland and Baltimore and Maryland, while changing the comma between Maryland and Baltimore to a semicolon to define these as two separate keywords.
Delete Headers

After the keywords all in order the colored headers are deleted and the file is saved a text-delimited file
S-40—Text File

The text delimited file has to be opened and all quotes found and replaced with nothing, and any extra space deleted from the end. The file and the .pdf are then uploaded into ContentDM. Departments are added from a list, and the correct thumbnail is selected. They are then uploaded to the ContentDM server, approved, and indexed, but we have substantive difficulties with the server and the approval and indexing processes.
Then we load the .pdf’s and tab delimited metadata into the contentDM client. We set 2 values in the client, the thumbnail and the UMBC department from an authoritative list. Finally, we load the files to the ContentDM serve and approve and index them.
What I’ve Learned
This project has been a substantial learning opportunity for me. I’ve gained a lot new Excel skills as well new skills in manipulating .pdf’s. I’ve also learned how to use ContentDM. Additionally I found that a project may entail much than is apparent at first and that when accepting work utilizing software that I’m unfamiliar with, it would be a good idea to investigate the quirks of that software and whether there will be sufficient support to troubleshoots problems.
Questions?