Migration & Ingest from CONTENTdm: Notes on Processing for Partners

As with all migration and ingest processing, we welcome discussion on any aspects because processes can always be improved for partner ease. Please let us know on any questions, concerns, ideas, etc. Please let us know on both direct concerns and thoughts overall because dLOC (as well as other collaborative projects for which UF is the technical host, other institutions and projects using SobekCM) benefit greatly from partner insights.

Overall

Migrating and/or ingesting files from CONTENTdm includes three areas:

- Collection(s) and subcollection(s) setup
- Metadata
- Files

Each of these is detailed below.

Subcollections

For existing collections and subcollections, partners should confirm that these should be created in the version in dLOC/SobekCM. This is an opportunity for changes, if wanted. Research confirms that subcollections are of great interest, value, and use for all users (advanced scholars, students, etc.). We recommend retaining/mirroring any existing subcollections that are already defined. In SobekCM/dLOC, subcollections have features at that curated level and separate landing pages to provide context for the materials, and also retain all rich support from the higher Partner Collection with all partner materials still directly accessible and supported through the main Partner Collection page.

For subcollections, one example is FIU's Collections in dLOC (http://dloc.com/ifiu). FIU's main collection includes all FIU items, and subsets of those items are also organized into two subcollections. These subcollections are linked from the bottom of the main FIU collection page (direct links: http://dloc.com/ifiulaw and http://dloc.com/ifiudloc). Having the subcollections for thematic/topical collections means that they can be more easily promoted as collections. For instance, all dLOC thematic/topical collections are listed on the thematic collections page: http://dloc.com/dloc1/collect

Process for Subcollections:

1. Partners confirm any/all subcollections to be retained (noting any changes for collection names, descriptions, etc.)
2. Partners confirm if the same banner from the Partner Collection page should be used for subcollections
    a. Or, partners provide an updated banner now, or can update using SobekCM's online tools
    b. Banners should be 900 pixels wide by under 150 pixels tall
3. Partners confirm that the text from existing collection pages should be used, provide text, or update the text using SobekCM's online tools

Process for Ingest of Items/Materials, including Files and Metadata

As of November 2013, the current ingest process from CONTENTdm[1] includes partners going through these steps:

1. Partner copies/exports files and metadata from CONTENTdm
    a. Partner makes a full copy of the CONTENTdm collection folders. The copied files are used for the next preparation steps, which may be best done by the partner of by dLOC's Technical Host (UF).
    b. Export the metadata using the Exporting to Tab-delimited Text Files method (includes the text of the pages and links to each of the files associated with each object).
2. File Processing by Partner *OR* dLOC's UF folks:
    a. In the copied folders only (with these version the version to send for ingest), go into each collection folder and delete all the subfolders EXCEPT *image* and *supp*.
    b. Delete all the small "icon" jpeg images
    c. Pull all the TIFFs, XML, and CPD files out of the *image* subfolder and into the root collection folder
    d. Use Adobe Photoshop batching with actions to convert any remaining JPEG images to TIFF (to allow the SobekCM builder to create its own derivatives)
    e. Once this work is complete and confirmed, move the new TIFFs into the collection folder and delete the remaining *image*subfolder
3. Metadata processing and full ingest by dLOC's UF folks continues with current process.

[1] Technical documentation and process notes, ongoing updates: http://dloc.com/sobekcm/migration/contentdm

Technical Documentation (copied Nov. 2013 from http://dloc.com/sobekcm/migration/contentdm)

MIGRATING FROM CONTENTDM™

RESOURCE TYPES

If you are migrating single image files from ContentDM to SobekCM, you should be able to use the Spreadsheet importer to easily create the new resources within SobekCM. Then, you need only write a small script to move the images into folders named with the new BibID_VID and drop those into the SobekCM Builder.

This gets somewhat more complicated when working with complex multi-page documents and supplementary materials.

MIGRATION NOTES

What follows are notes regarding a migration from ContentDM to SobekCM in October of 2013. These notes are posted in the hopes that this will make future migrations simpler. Having never had a collection in ContentDM, there is a very good chance that there may be a better way. If you know of anything to make this process easier, please do not hesitate to contact me at Mark.V.Sullivan at Gmail.com.

Preparing the collection folders for import

1. I received an exact copy of the collection folders from ContentDM. If you are not already working a copy of your collection folders, make a copy now.
2. Step into each collection folder and delete all the subfolders EXCEPT *image* and *supp*.
3. Delete all the small "icon" jpeg images
4. Pull all the TIFFs, XML, and CPD files out of the *image* subfolder and into the root collection folder
5. Use Adobe Photoshop batching with actions to convert any remaining JPEG images to TIFF (to allow the SobekCM builder to create its own derivatives)

6. Once this work is complete and confirmed, move the new TIFFs into the collection folder and delete the remaining *image*subfolder
7. I then moved all the prepped source folders into a new folder for processing below

All of this work listed above was done by hand, although it would be very simple to automate much of this with simple scripting.

Preparing the collection-level metadata for processing

1. I received collection-level metadata output from ContentDM, which included the text of the pages and links to each of the files associated with each object. This metadata was exported using the *Exporting to Tab-delimited Text Files* method. One of the most interesting things about this format is that each individual page for a complex, multi-page object is listed AND the multi-page complex object is also referenced, in the same file.
2. Convert each individual collection output (txt) into Excel for ease of working with them
3. Add a new column at the beginning of each Excel file with the new SobekCM collection code
4. Combine all of the separate collection-level spreadsheets into a single spreadsheet for processing everything at the same time
5. Add a new column at zero position named ID and fill with series starting at 1, 2, 3, etc..
6. For rows that are multiple issues of the same title (in a newspaper or periodical) set the ID to be identical

Process the files and metadata

1. Using code included here, check that all the files exist (see *Verify_Resource_Files_Exist()* within code)
2. Create text files from the text in the spreadsheet (see *Add_Text_Files()* within code)
3. Step through and process any referenced CPD files. Move the CPD file and all related images and text into their own subfolder for processing as a single item. (see *Process_CPD_Files()* within code). Note: this implies that the next time we go through the spreadsheet, when we find a row that references a page within a CPD file, it will not be found. This is why we checked that all files existed at the beginning of this process.
4. Finally, build the complete METS packages from the spreadsheet, CPD folders, and loose files (see*Create_SobekCM_METS()* within code below)

C# CODE

The code below essentially follows the steps listed above for the final processing of the metadata and images.

```
// Read the prepared Excel spreadsheet into a DataTable
ExcelBibliographicReader xlsReader = new ExcelBibliographicReader();
xlsReader.Filename = "Complete.xls";
xlsReader.Sheet = xlsReader.GetExcelSheetNames("Complete.xlsx")[0];
DataTable importTbl = xlsReader.Check_Source();

// Check that all files exist

ContentDM_Importer contentDm = new ContentDM_Importer(importTbl, @"\\ad.ufl.edu\....\
College\source");
contentDm.Verify_Resource_Files_Exist();
Console.WriteLine();

// Since the text is in the spreadsheet, write out the text files for
// indexing within Sobek
int text_files_written = contentDm.Add_Text_Files();
Console.WriteLine("Wrote " + text_files_written + " text files");
Console.WriteLine();

// Process all the CPD files referenced
int cpd_files_handled = contentDm.Process_CPD_Files();
Console.WriteLine(cpd_files_handled + " CPD files handled");
Console.WriteLine();

// Create the METS packages ready for SobekCM
contentDm.Create_SobekCM_METS(@"\\ad.ufl.edu\....\College\ready\");

Console.WriteLine("COMPLETE");
Console.ReadLine();
```

This code uses the classes found in the ZIP file below, as well as the *SobekCM_Resource_Object* library, which is available in the SobekCM source code from our GitHub site.

Download ContentDM_Importer C# class.

TRADEMARKS