
THE OHIO STATE UNIVERSITY
MOMMY: MASTER OBJECTS METADATA AND MIGRATION—YEAH!
go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

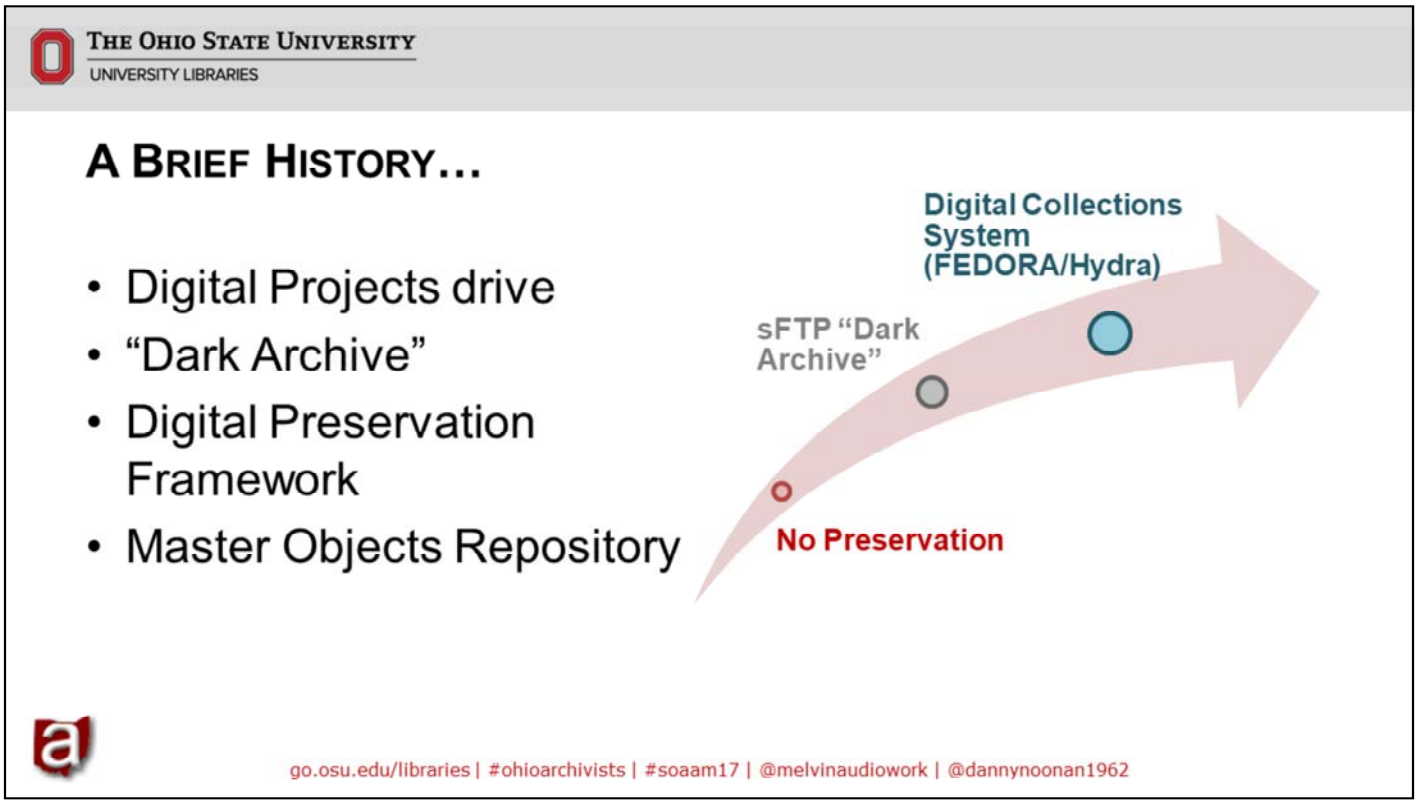
THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

WHERE ARE WE GOING?

- Setting the stage: A very brief history…
- Project planning & migration preparation
- Identification, transformation, normalization of metadata
- Challenges for our further adventures…

go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

Good afternoon! Darnelle and I will be leading you on a brisk trip, exploring our adventures in migration—from organized chaos to a FEDORA/Hydra-based preservation environment. I will set the stage with a brief history of our digital preservation efforts and then provide an overview of our project planning and migration prep activities. Darnelle will then navigate us through the seas of identifying, transforming and normalizing our metadata prior to ingest. Lastly, we will identify our existing challenges as our migration activities move forward under full steam.



So a long time ago in a library far, far away...a story whose origins are lost in the mists of time (or at least more than a decade ago). No one can say definitively that this is what happened, but it is what I have been able to piece together, along with my own first-hand knowledge over the past ten years. Outside of a couple of big projects dedicated to brittle books and these and dissertations, the bulk of digitization efforts were conducted by our Special Collections and Archives personnel. As they began to fill departmental share-drive space with their projects, the Libraries began to run out of digital storage space. To accommodate this growing digital mass, the Libraries IT department pulled an old web server out of mothballs to create a shared drive known simply as K.

Sometime around a decade ago that drive became unstable and its contents were unceremoniously placed on a server known as dspace04. This gave rise to the myth that their collections were being preserved in DSpace. NO, dspace04 was just a sever used for staging DSpace upgrades and ingests, and just happened to have excess capacity. With the dumping of the K-drive server space contents onto dspace04 (and not into DSpace itself), we ended up with all sorts of digital materials this server. We had inadvertently created the "Dark Archive" with little or no consideration for what we were putting there—not only did we get digital master objects, but derivatives, working files and various detritus.

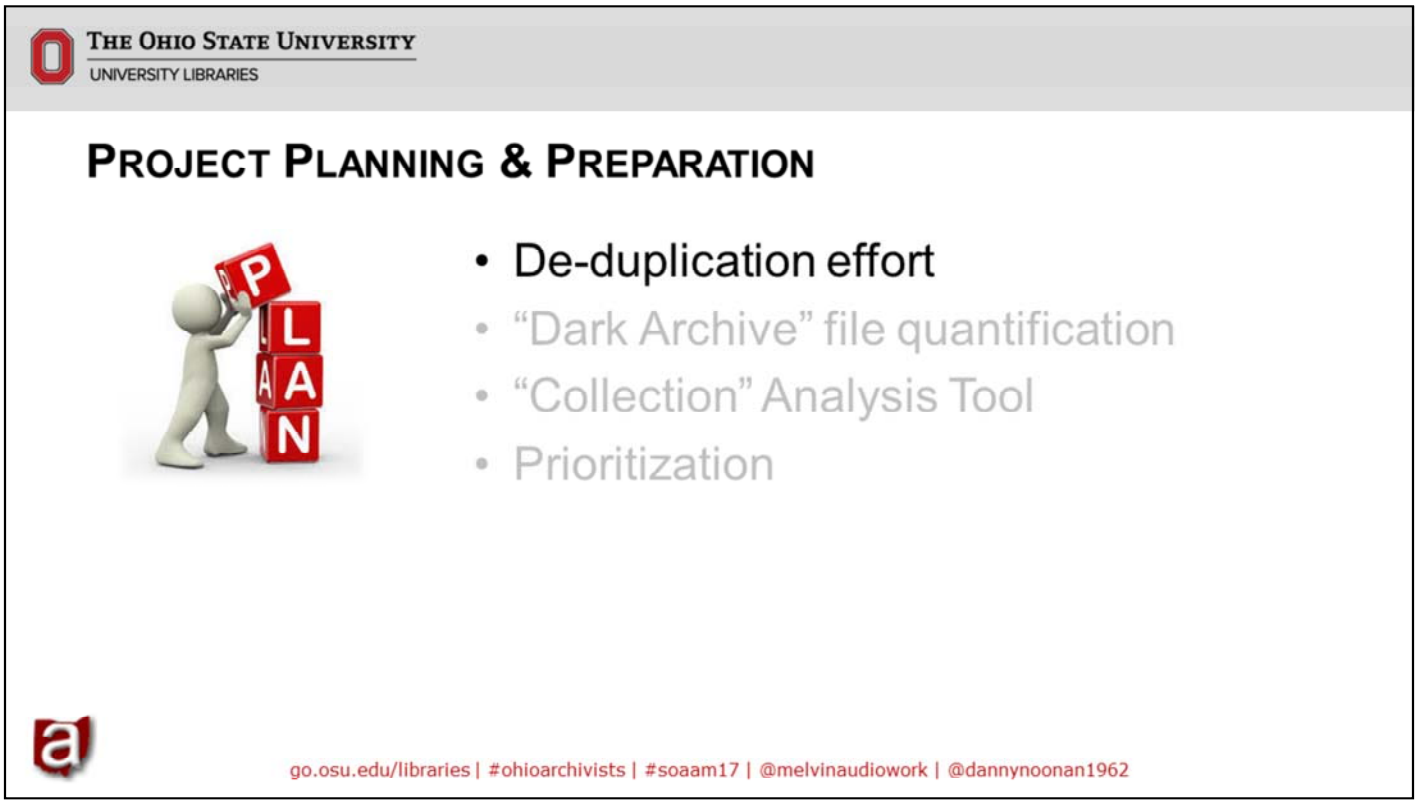
However, a key benefit of the Dark Archive is that it did/does provide controlled access through sFTP. As such, the Libraries did subsequently begin to take more prescribed steps in deciding what it put onto this server, while carving out a new K-drive space to actively work on projects; however, this did not curtail excessive amounts of duplication of master and derivative objects, nor was there any official policy around its use; and lastly good file management policies/techniques/processes were not used nor in many cases basic metadata created.

In 2012 and 2013 a team from the OSU Libraries participated in the DigCCur Institute. Our project was the development of a digital preservation policy framework that began to set the stage for migration to a true preservation environment. This effort dovetailed with the hiring of our Head of Digital Initiatives, Terry Reese, who is the chief architect of our new FEDORA/Hydra preservation environment. In 2014 he spearheaded the Master Objects Repository Task Force, which laid out a framework for our digital preservation activities including:

- defining Master and Derivative Objects
- defining the environment and high-level management processes for a Master Objects Repository (MOR) within the Libraries' digital storage environment
- recommending procedures for proper deposit and registration of appropriate objects in the MOR including workflows and metadata for management/identification purposes, including interactions with other systems as appropriate.

The recommendations were software/hardware agnostic to allow digital Master Objects to be migrated to and preserved on future storage platforms.

Subsequently in 2015 the Libraries decided to implement a FEDORA repository solution using Hydra-based Sufia for our user interface.

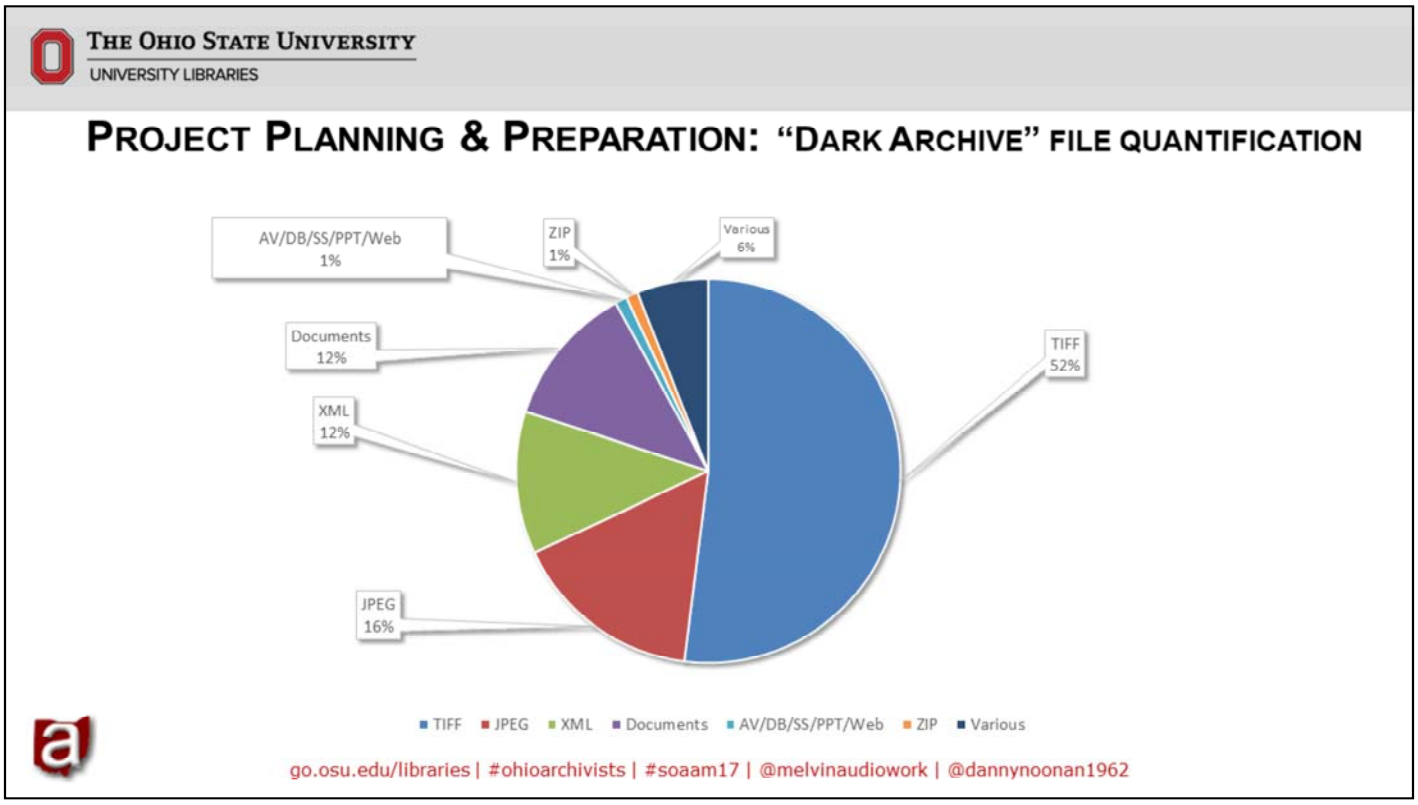


So where to start. As early as late 2011, the Libraries engaged a retired librarian to conduct a rudimentary inventory of our digital stuff. I inherited this inventory that covered not only items in our Dark Archive, but also our DSpace repository know as items KnowledgeBank, our shared drives and items on loose media. Through some educated interpretation and SWAGging I estimated that we had upwards of 14TBs that likely needed to find a home in a true preservation environment. One of the things this inventory lacked was a comprehensive look at our Dark Archive and its contents. An oh yeah, the libraries had put another K-drive in place to ostensibly work on digitization projects.

As we began to examine the Dark Archive, one thing we were certain of was that there was/is a significant amount of duplication within the it and with the replacement K-drive and the departmental/committee shared J-drive. In conjunction with the development the digital preservation policy framework, we started to conduct a de-duplication effort on the Dark Archive, where we identified over 215,000 duplicates. This was driven by the fact that we were running out of digital storage space at the time. Working with our IT Infrastructure Support group, we developed spreadsheets that identified file-paths for duplicate pairs (and sometimes triplicate, quadruplicate or more). In sharing these with the responsible collection archivist or curator, we discovered that they also likely had copies on the K-Drive. So we did a Dark Archive vs K-drive analysis with the intention of retiring the use of the K-drive and making certain all masters were in the Dark Archive and derivatives

distributed to their appropriate access point. By mid-2014 we had made significant headway of de-duping the Dark Archive and had finally retired the K-drive (or so we thought, but more on that at the end).

2015 saw the implementation of our FEDORA/Sufia platform whose pilot content was Libraries' collections content migrated from an external system that another campus entity no longer supported.

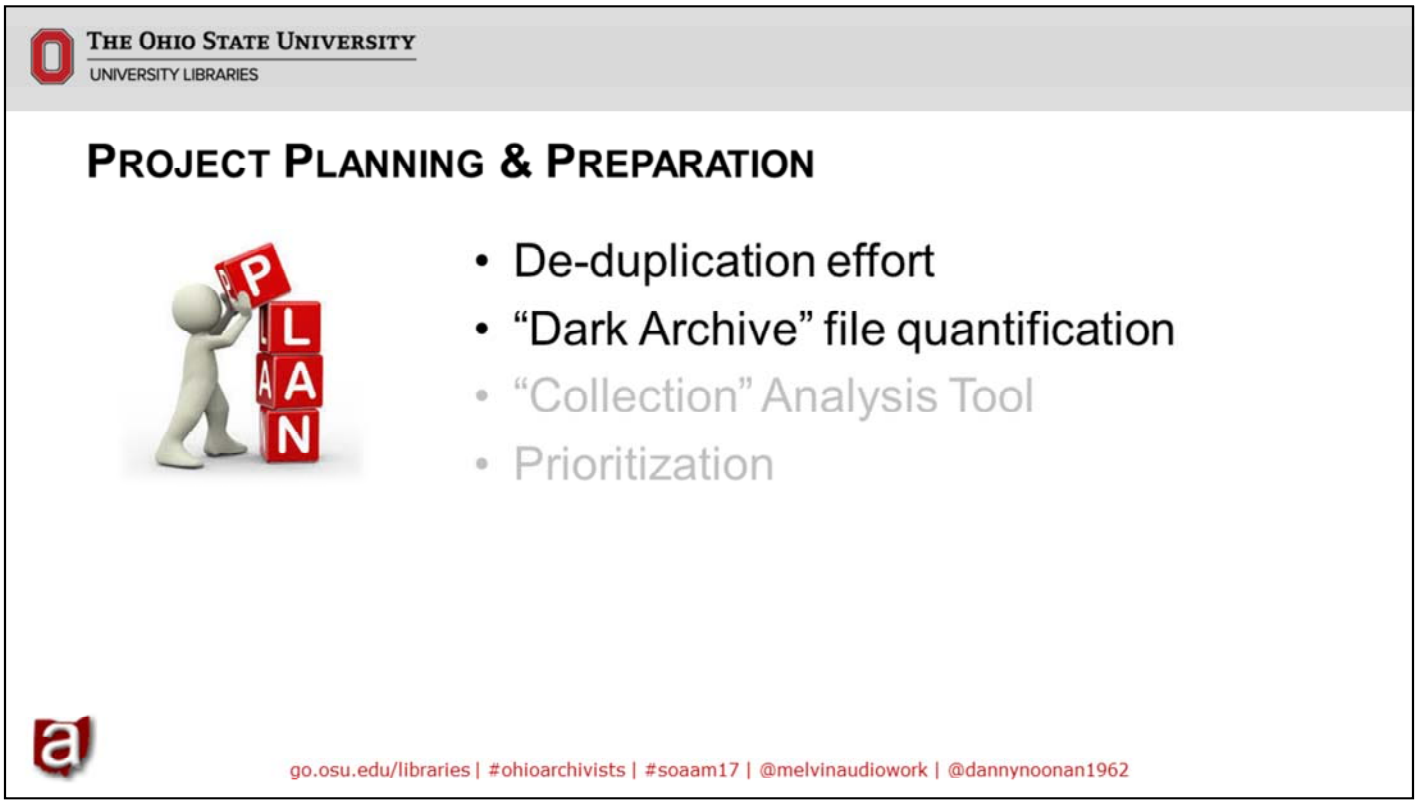


In preparation for migration of content from the Dark Archive, we identified more than 85 files types and nearly 2,000,000 objects that needed to be considered for migration. The good news was that 52% were TIF images that for the most part should be a no-brainer for migration. The next largest quantity of files were JPEGs which may be masters or derivatives; documents, the bulk of which are PDFs; XML which may be metadata, but the bulk of which are poorly formed faux-xml; various AV, DBs, Spreadsheets, PPTs and web-files; and zip files, whose internal contents will need to be examined.

There remaining 6% are obscure file types that may or may not need to be migrated or are the result of poor file naming practices.

We now knew how many things we had, but who do they belong to and how do we prioritize the migration of more than a million items?

And oh, what about all the metadata that will be needed, because one thing is absolutely clear: NOTHING GOES INTO THE MASTER OBJECTS REPOSITORY WITHOUT A MINIMUM AMOUNT OF METADATA!!!!!!

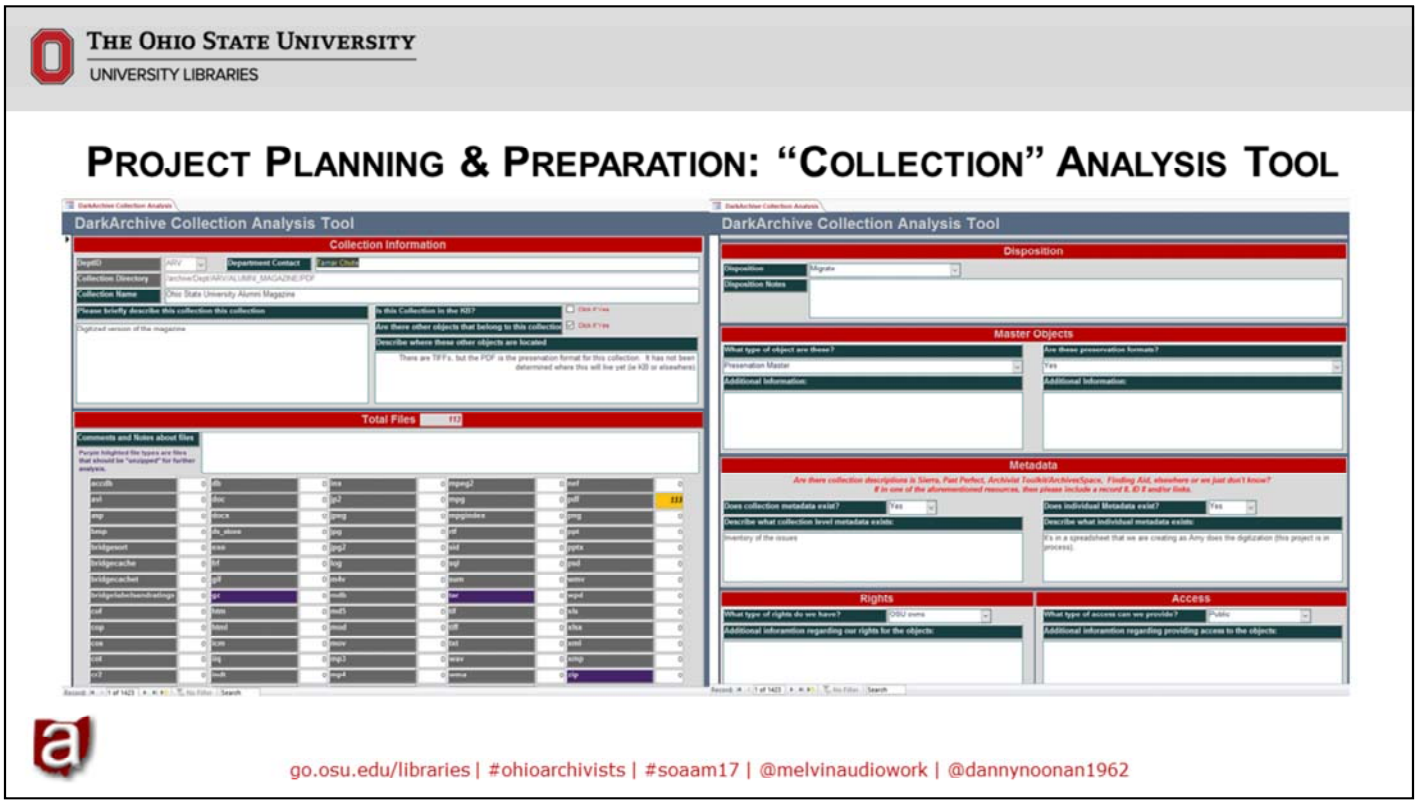


Now, before I turn it over to Darnelle to discuss metadata and workflow, let's look how did we approach the prioritization?

Fortunately, the Dark Archives' folder structure is set to coincide with collection owners. Right off the bat, we put 47% of the files on the back-burner as they are either master objects or support files for items in the KnowledgeBank—those will be the last files we will examine when we determine our strategy for interaction between DSpace and FEDORA.

Nearly a quarter of the files account for 11 collections within the University Archives, which are mostly from the Office of the President's document management system.

The remaining files—just shy of 30%—belong to 6 groups spread over approximately 150 collections, which means a lot more detailed analysis.

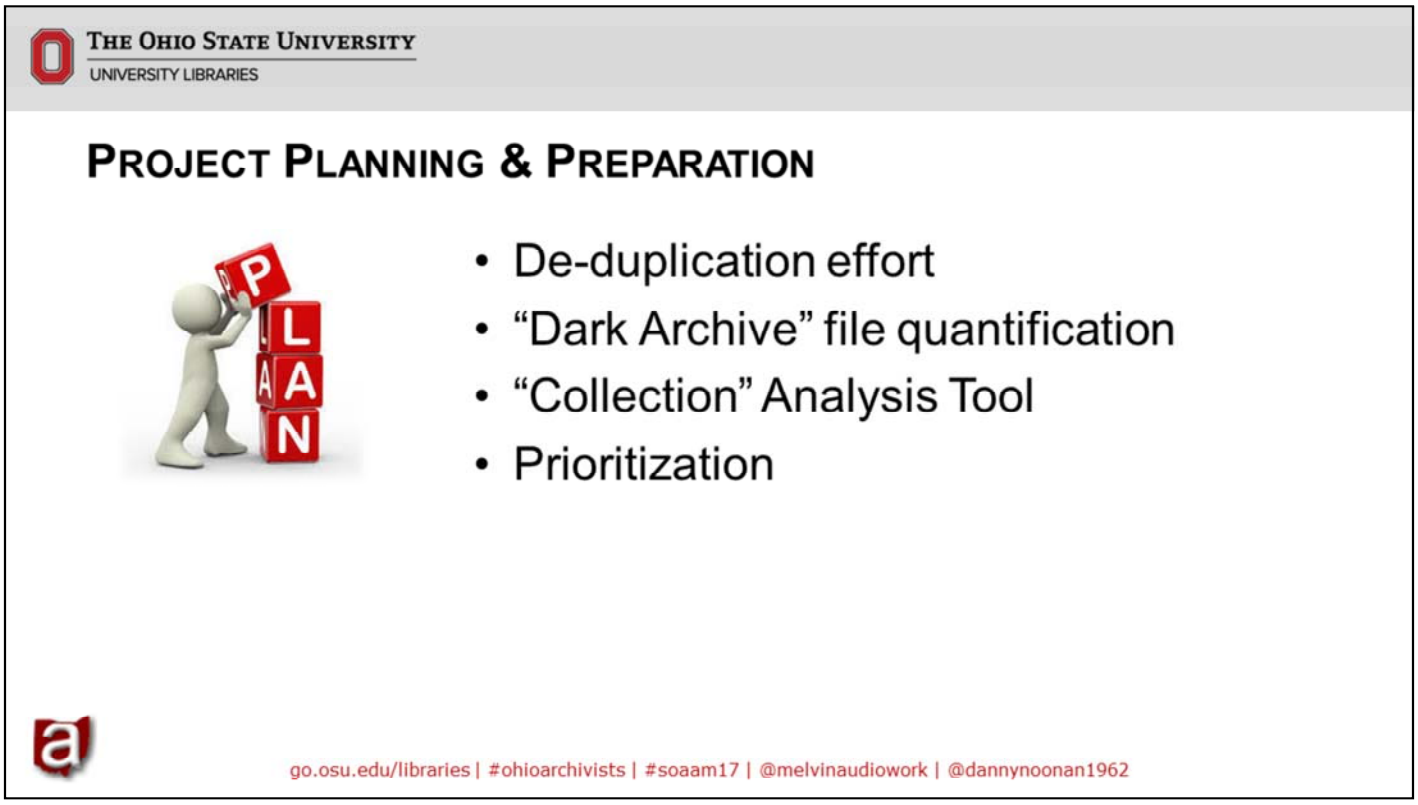


I constructed an Access database based upon the file-paths in the Dark Archive that was then shared with the appropriate archivists and curators that:

- Presented the file-path and quantification of file types within, and then
- asked them to Identify:
  - the collection the items belonged to?
  - Are there other objects that belong to this collection and where are they located?
  - Whether the objects should be migrated or disposed, or are need of further processing or assessment?
  - What type of object are these?
    - Preservation Master
    - Provisional Master
    - Working copy
    - Access Copy
    - Reproduction Copy
  - Are these preservation formats?
  - Does collection level and individual metadata exist and if so, where?
  - What are the intellectual property rights?
    - Public Domain
    - OSU Owned
    - Donor Owned

- Mixed
- Unknown

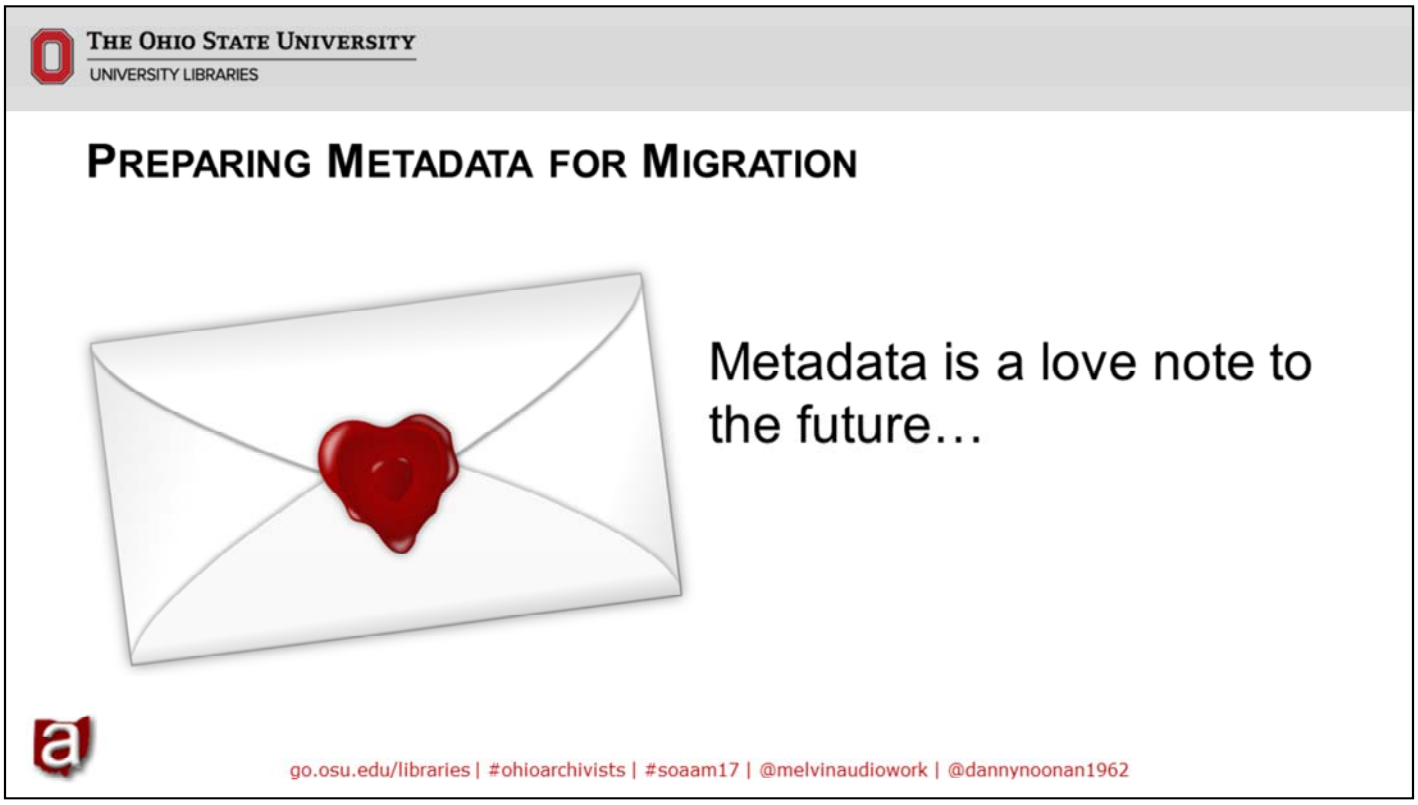
- And what type of access we are allowed to provide?
  - Public
  - Reading room
  - Private
  - Closed



This allowed us to begin to prioritize the migration, by identifying those collections with the most available metadata, and in which we can provide public access.

An additional wrinkle, that is shortly to go away, is that the only file profile our existing data model could handle was that of images, which are simple objects. Therefore, our highest priority were those collections that were TIFF or mostly TIFF that were of a single object nature. Secondary considerations were given to other image types, followed by documents, complex objects and audio-visual objects. With the upgrades to FEDORA and Sufia that are currently being deployed, we should shortly be able to accommodate ingest of other data models that included complex objects and other non-image file types.

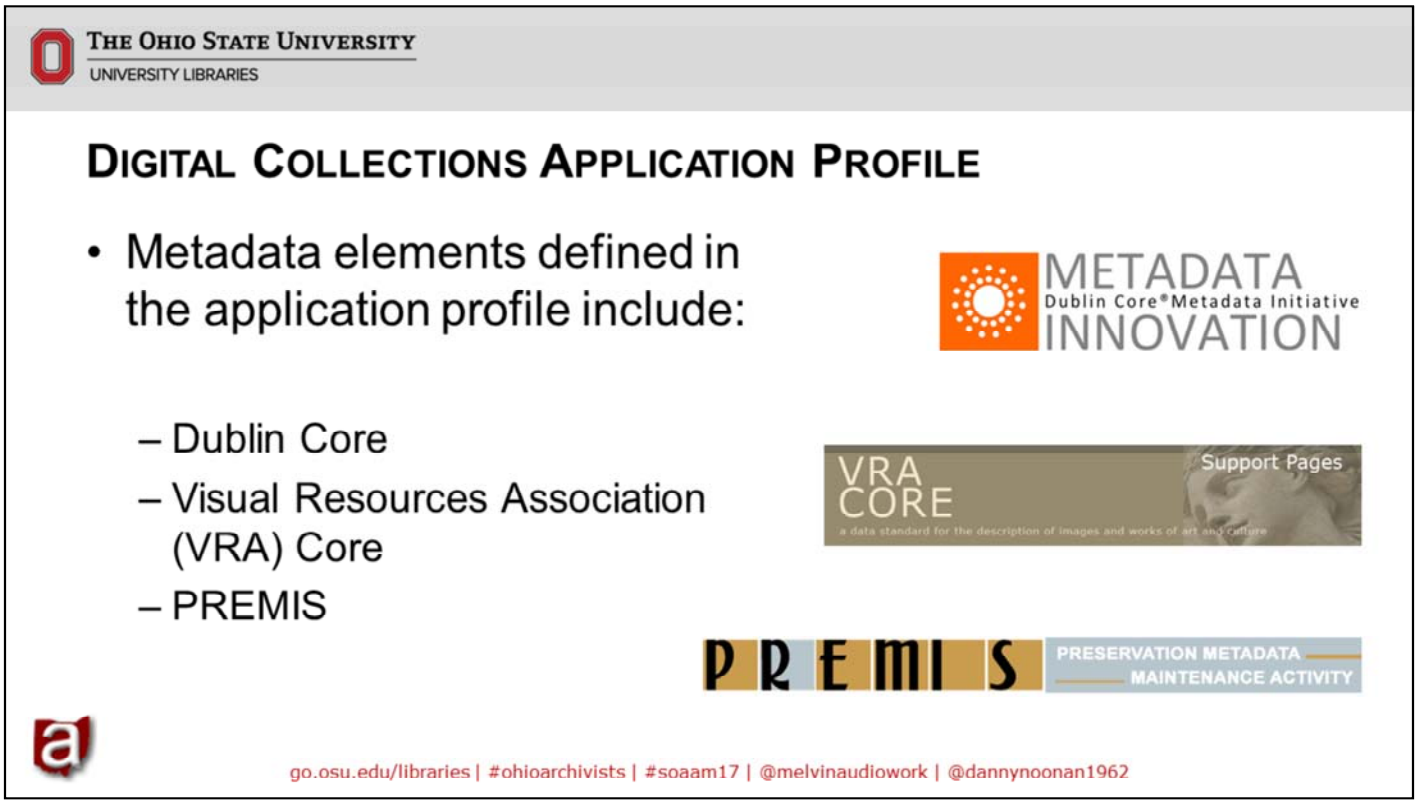
And with that I will turn it over to Darnelle and try to catch my breath.



Thanks Dan,

So as you all know, metadata can sometimes be messy, complex, incomplete, inconsistent, non-existent, and the quality can vary from resource to resource and collection to collection. Good metadata can increase discoverability, enhance the user experience, and identify the parties involved in the creation or contributions of a given work. Good metadata can also, act as an access point to the collection or resource once ingested into the repository.

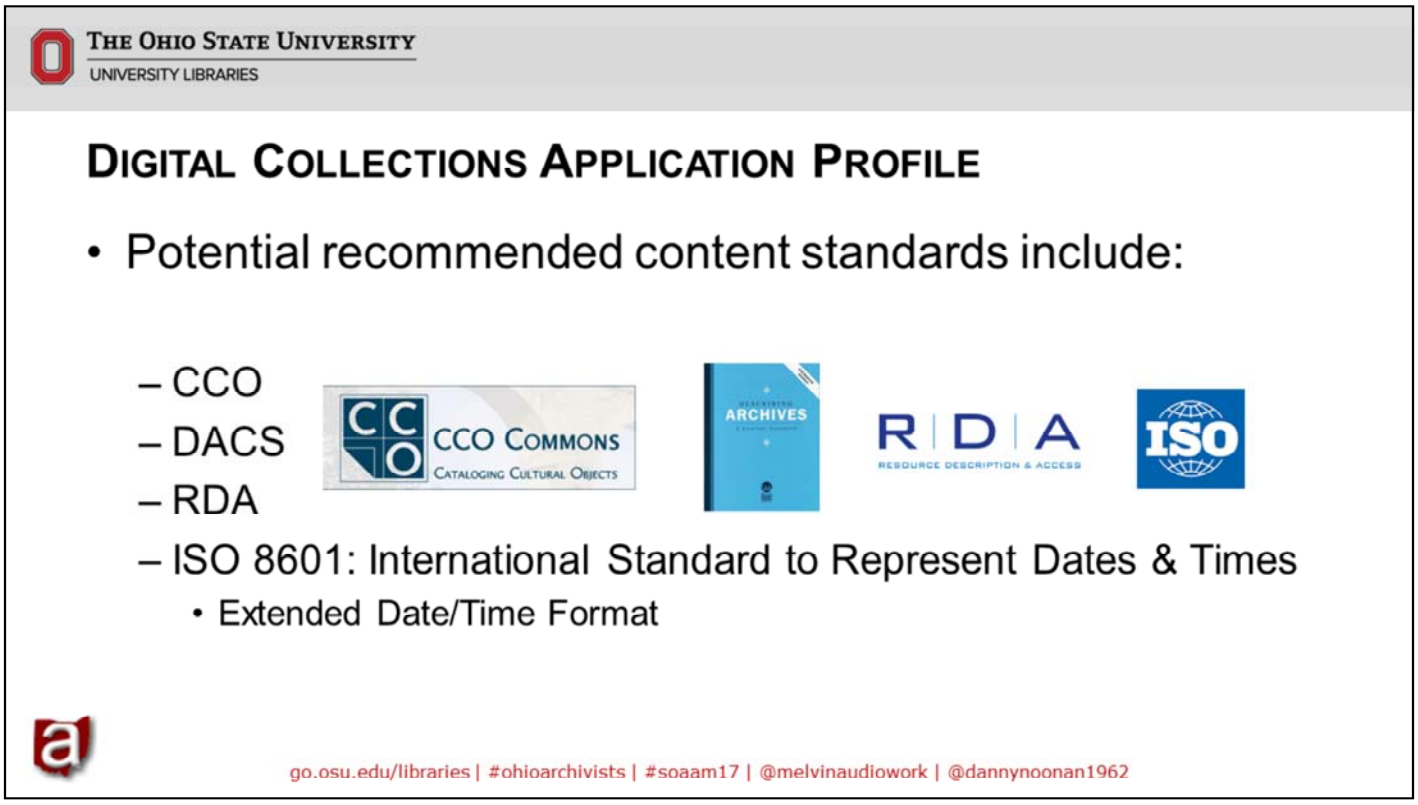
As the slide states, metadata truly is a love note to the future.



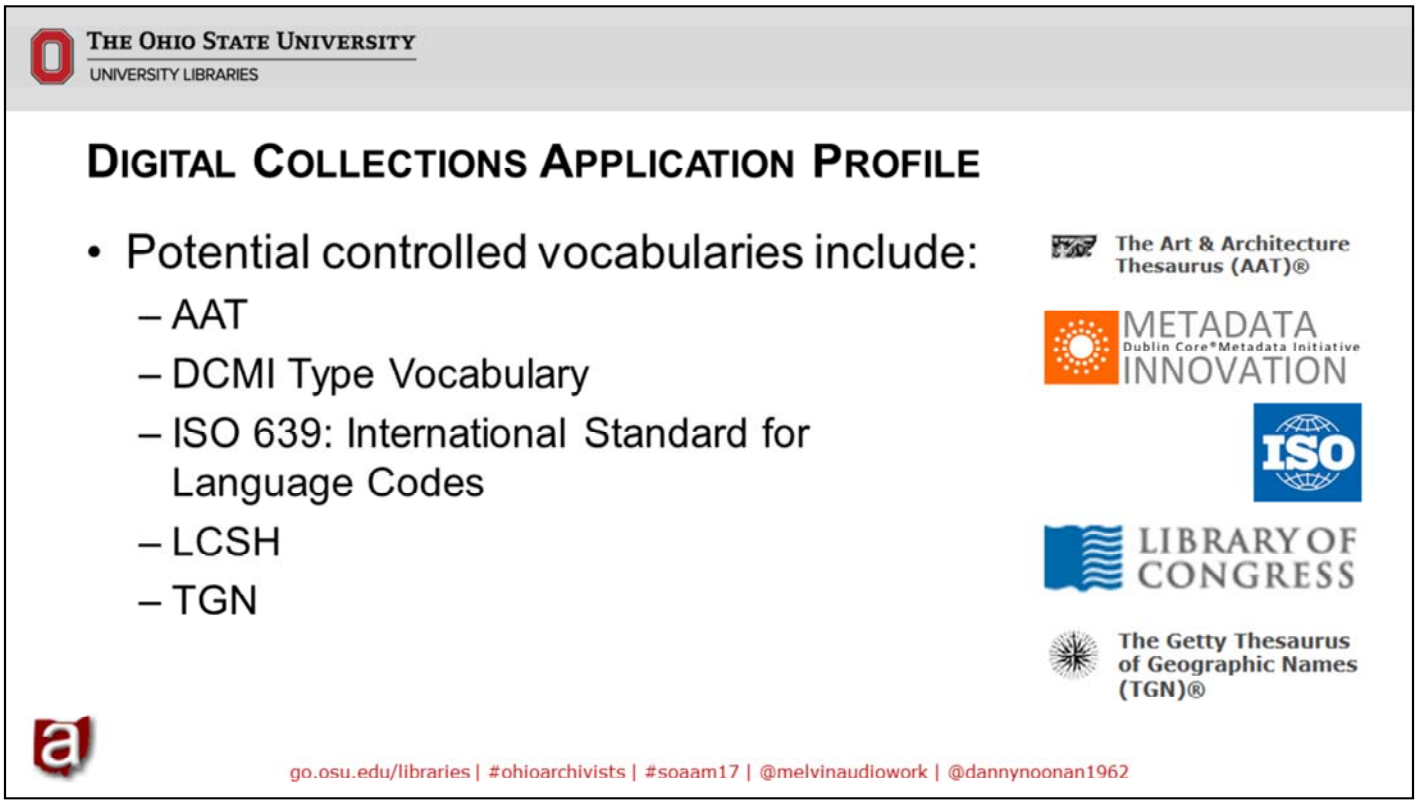
Before we began our work on this migration project, a lot of preparation was put forth to develop a blueprint for metadata mapping. This led to the creation of the Digital Collections Metadata Application Profile, which was an effort of University Libraries, Metadata Working Group.

The application profile is a group of metadata elements, attributes, and data values which includes policies, guidelines and examples for resources intended to be stored in the Master Objects Repository. The elements defined in the application profile were drawn from established metadata schemas such as: Dublin Core, VRA Core, and PREMIS.

In addition to the Application Profile, we recommended to our stakeholders to utilize established content standards and controlled vocabularies.

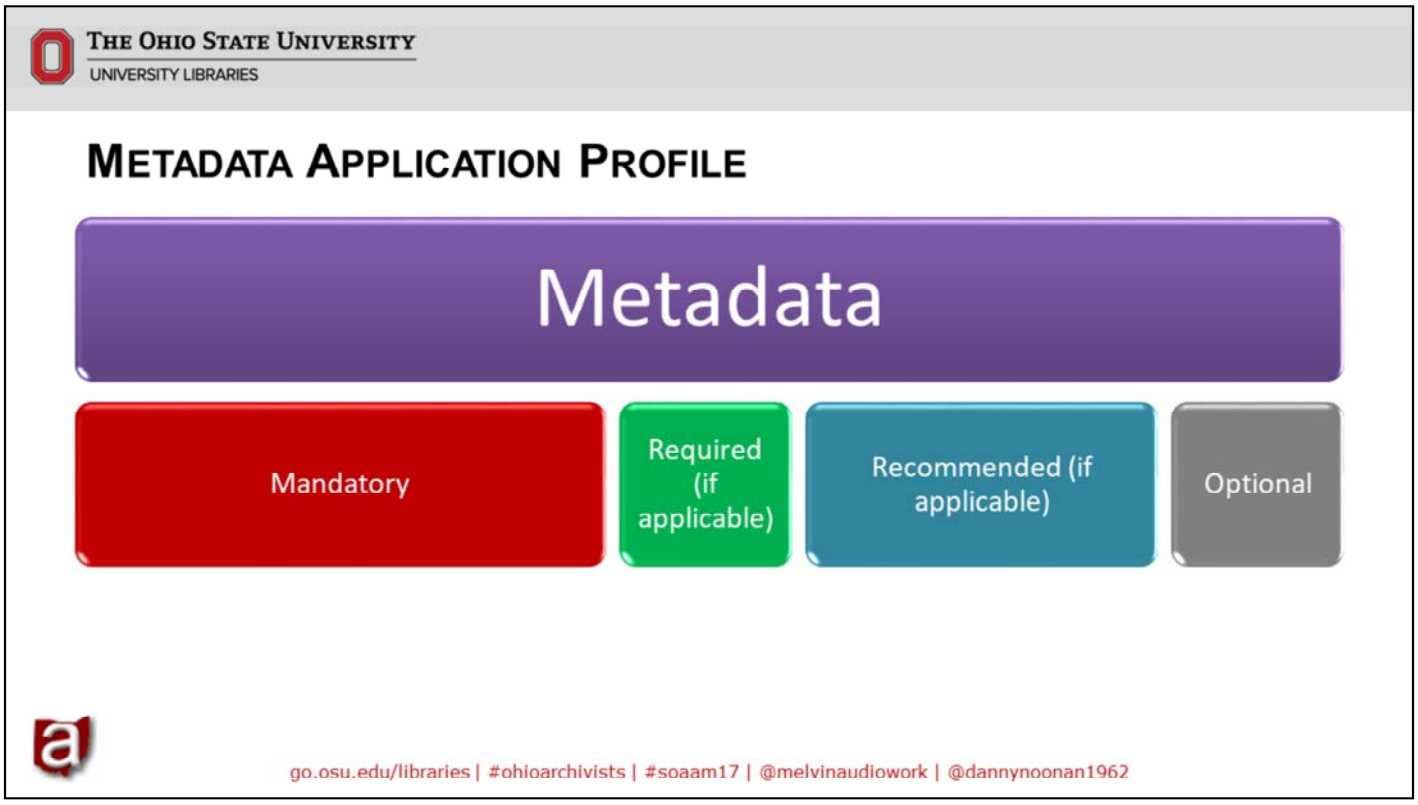


Examples of potential recommended content standards include: CCO, DACS, RDA, the International Standard to Represent Dates and Times, and Extended Date/Time Format.



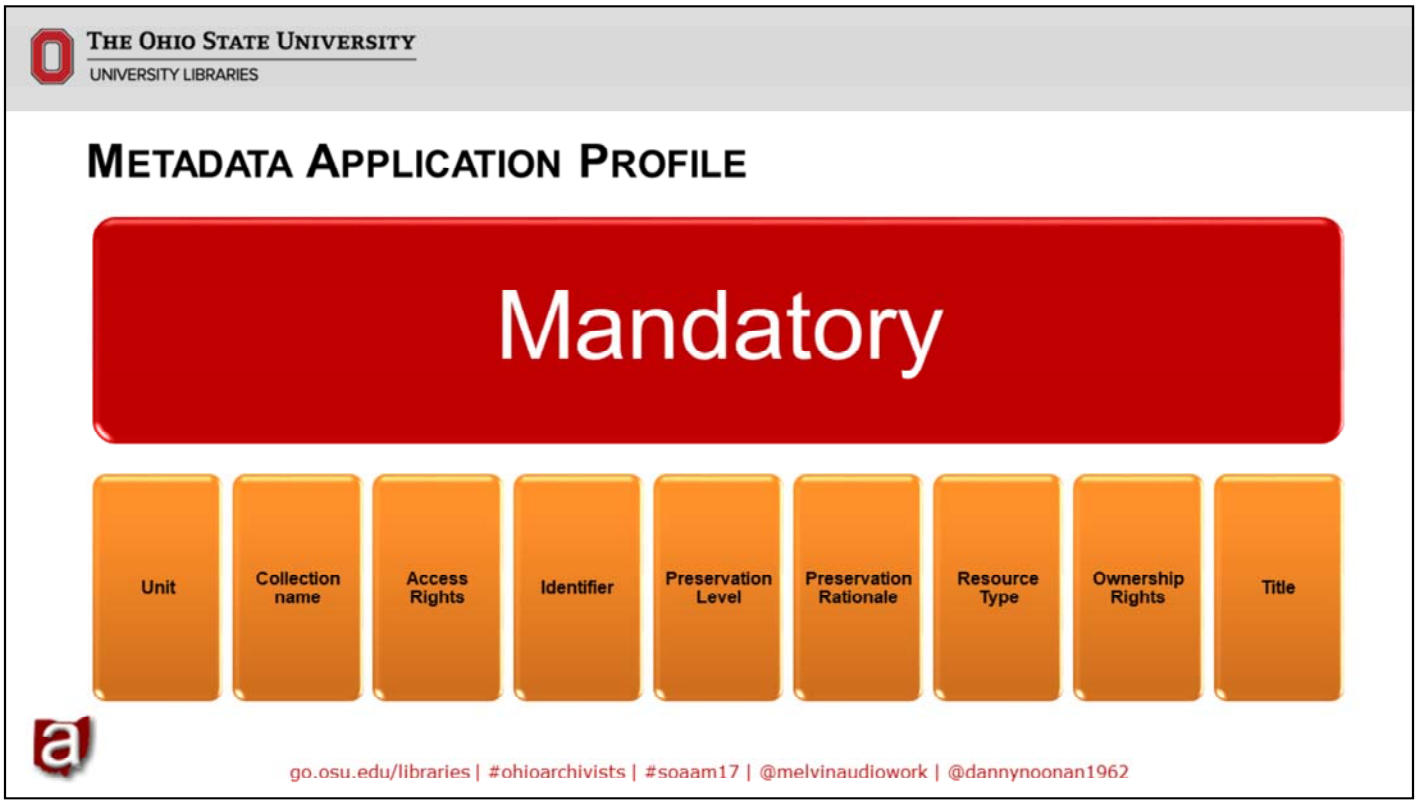
Examples of potential controlled vocabularies include: Art and Architecture Thesaurus, DCMI Type Vocabulary, International Standard for Language Codes, LCSH, the Getty Thesaurus of Geographic Names.

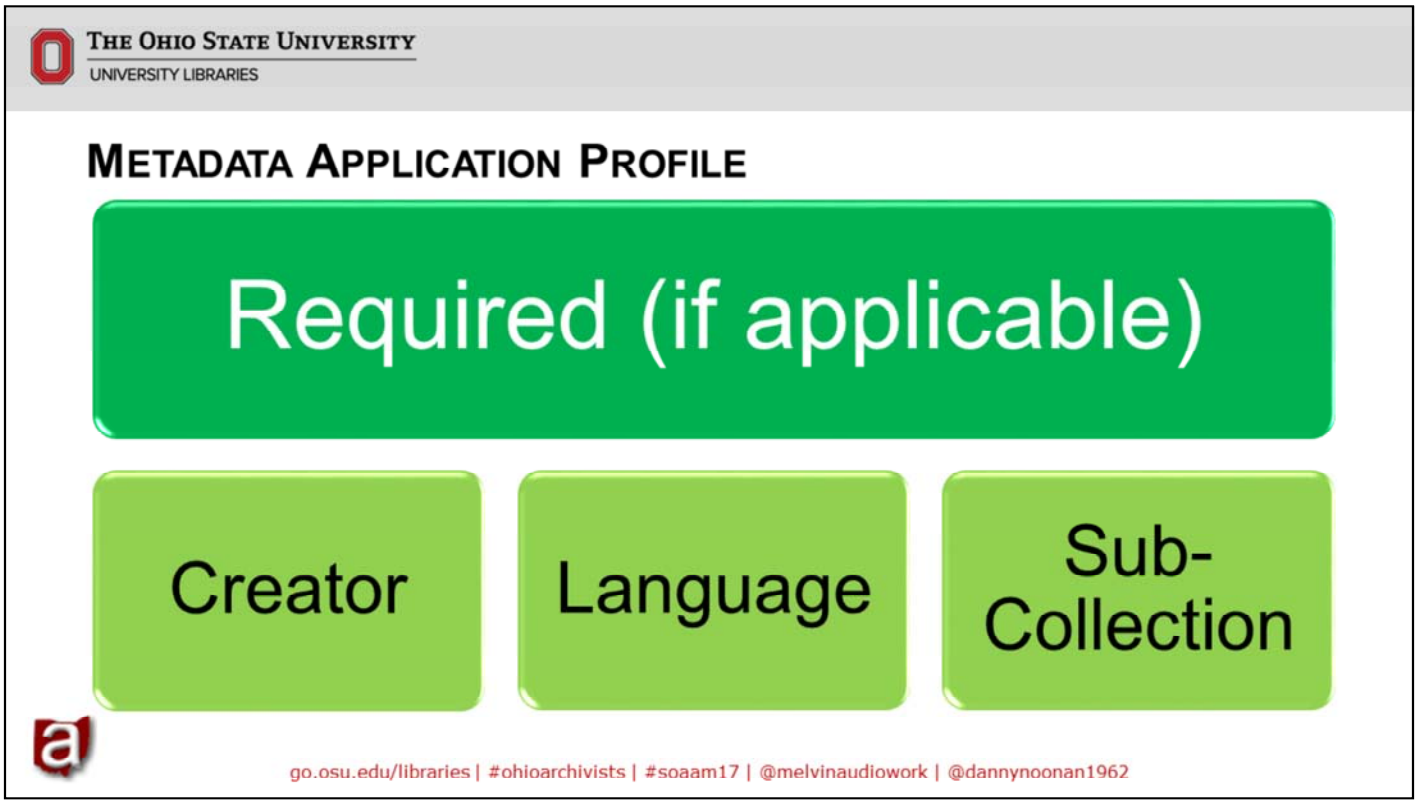
Since multiple special collections units are ingesting content into the Master Object Repository, we needed an application profile that was robust but also flexible for a wide variety of stakeholders and actions.

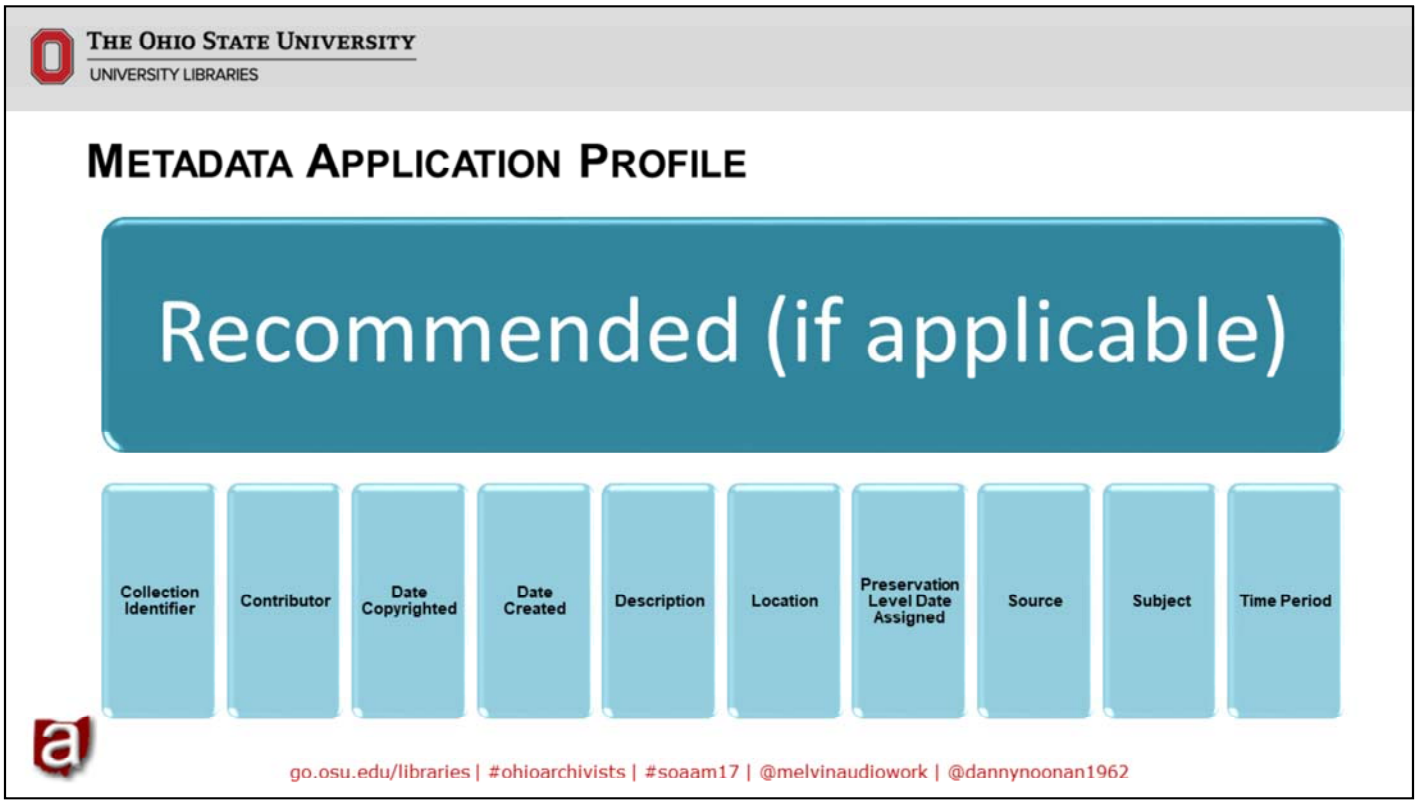


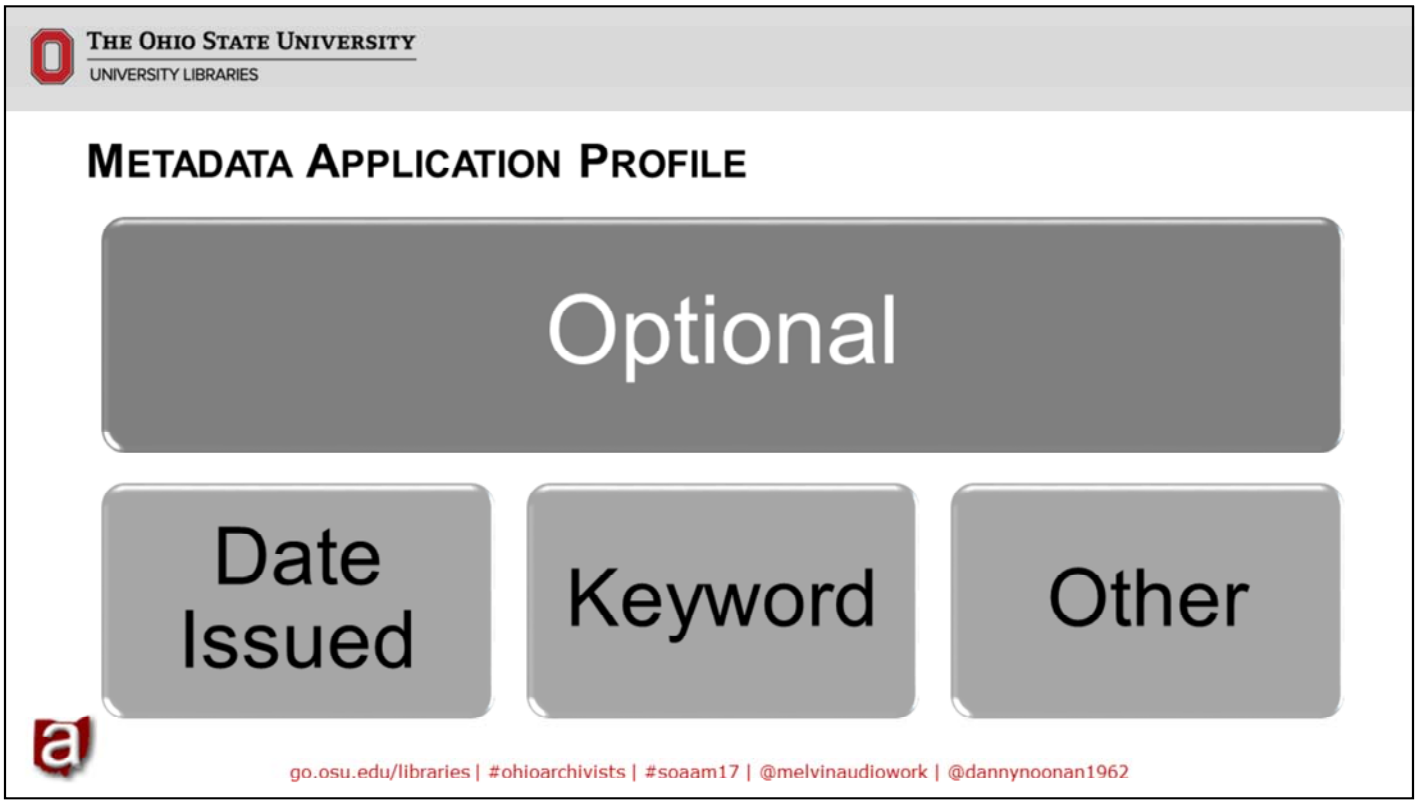
From this application profile, we established subset of elements that are the minimum required fields for any item to be ingested:

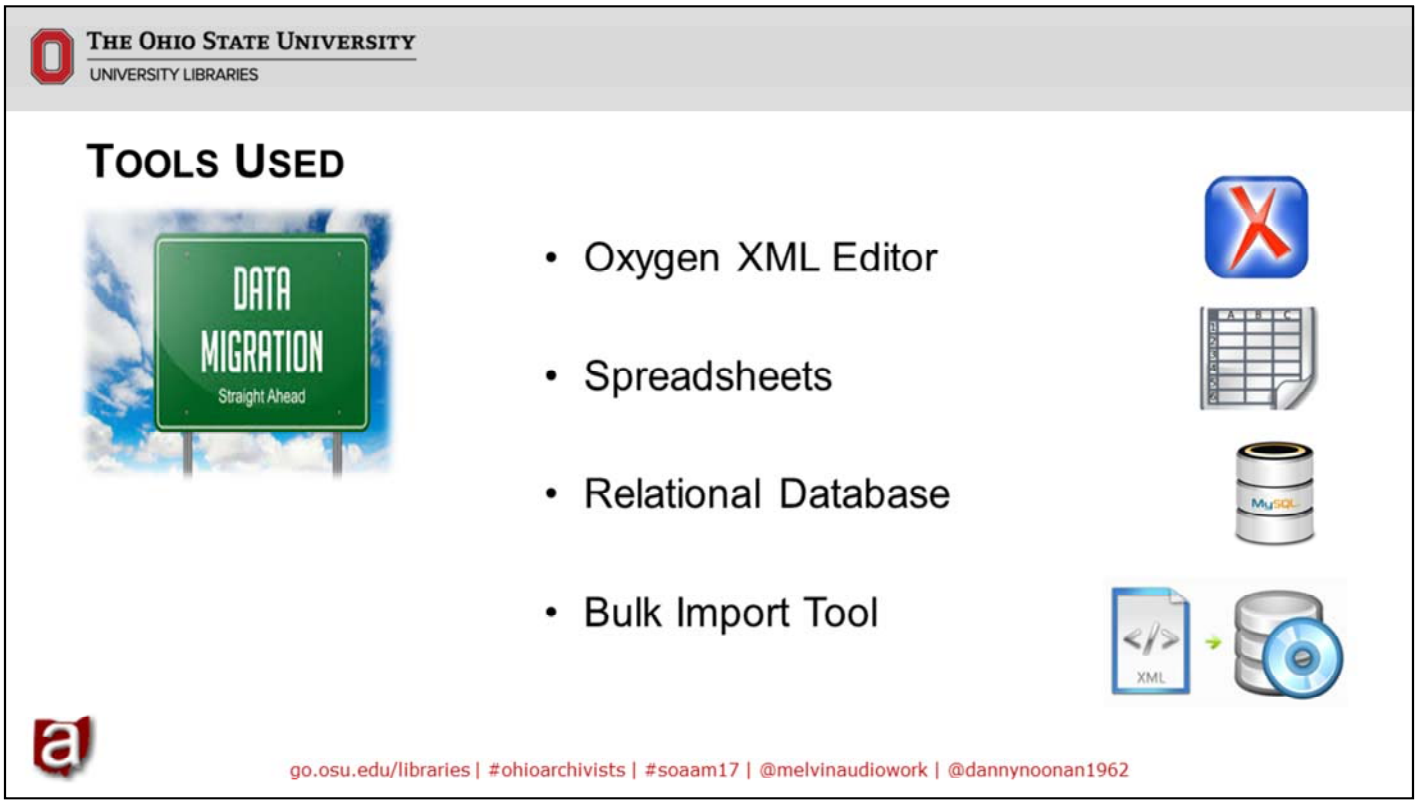
- Unit owner
- Collection name
- Access Rights
- A unique Identifier
- Preservation Level & Rationale
- Resource Type
- Ownership Rights
- Title
- And if applicable:
    - Creator
    - Language
    - Sub-collection


THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES
METADATA APPLICATION PROFILE
Mandatory
Unit
Collection name
Access Rights
Identifier
Preservation Level
Preservation Rationale
Resource Type
Ownership Rights
Title
go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962


THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES
METADATA APPLICATION PROFILE
Required (if applicable)
Creator
Language
Sub-Collection
go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

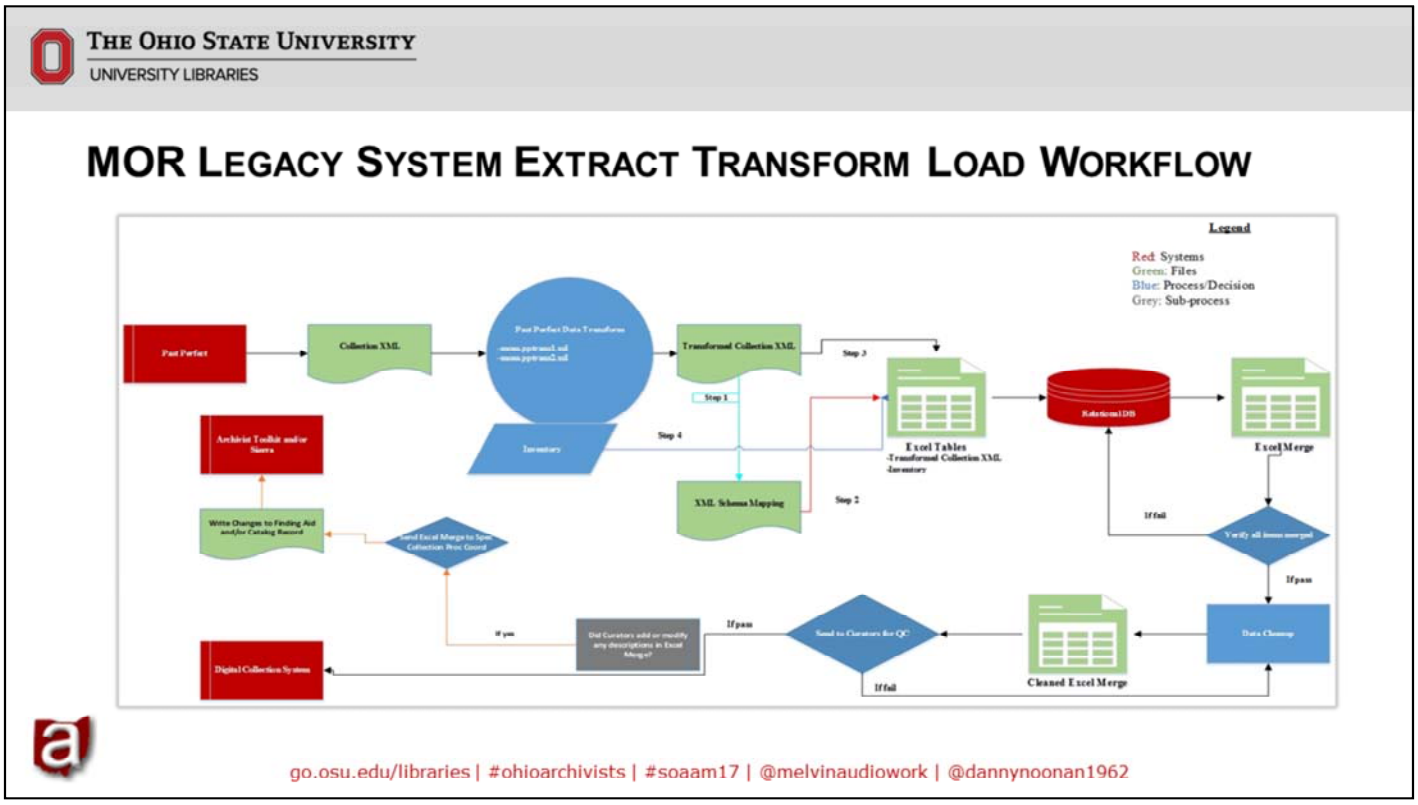

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES
METADATA APPLICATION PROFILE
Recommended (if applicable)
Collection Identifier
Contributor
Date Copyrighted
Date Created
Description
Location
Preservation Level Date Assigned
Source
Subject
Time Period
go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962


THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES
METADATA APPLICATION PROFILE
Optional
Date Issued
Keyword
Other
go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

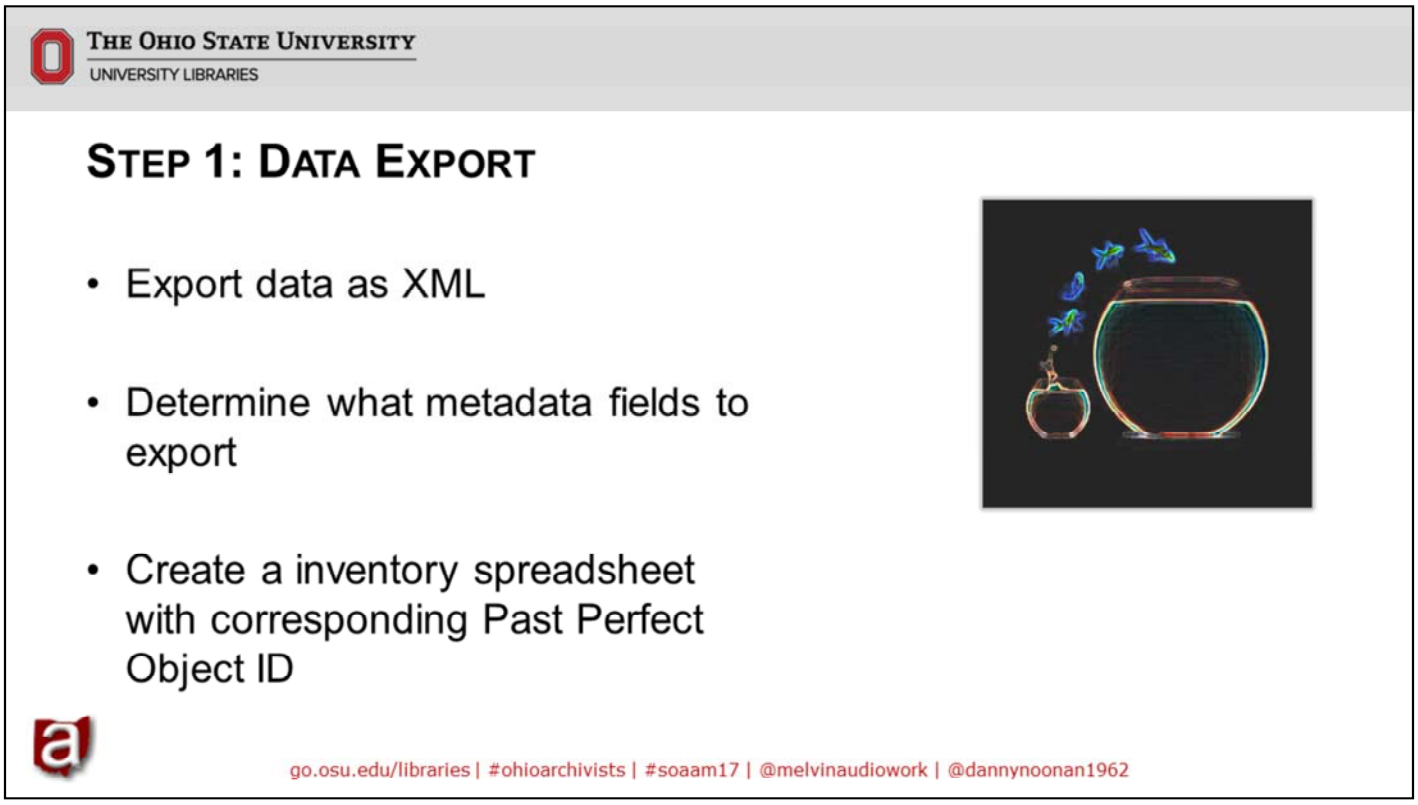


The four common tools used most for this project are the Oxygen XML Editor, Excel to generate spreadsheets, some type of relational database, such as Access or MySQL, and a locally developed Bulk Import Tool designed to ingest simple and complex objects into the Master Objects Repository.

So let's turn to discuss two workflows. The first workflow follows the Extract, Transform, Load process or ETL process. The second workflow was developed for the ingest of government documents that may have little metadata. In this instance, metadata is extracted from the file, and repurposed.



So to look how these steps fit together, here is a flowchart that illustrates this workflow.



Step 1, Data Export.

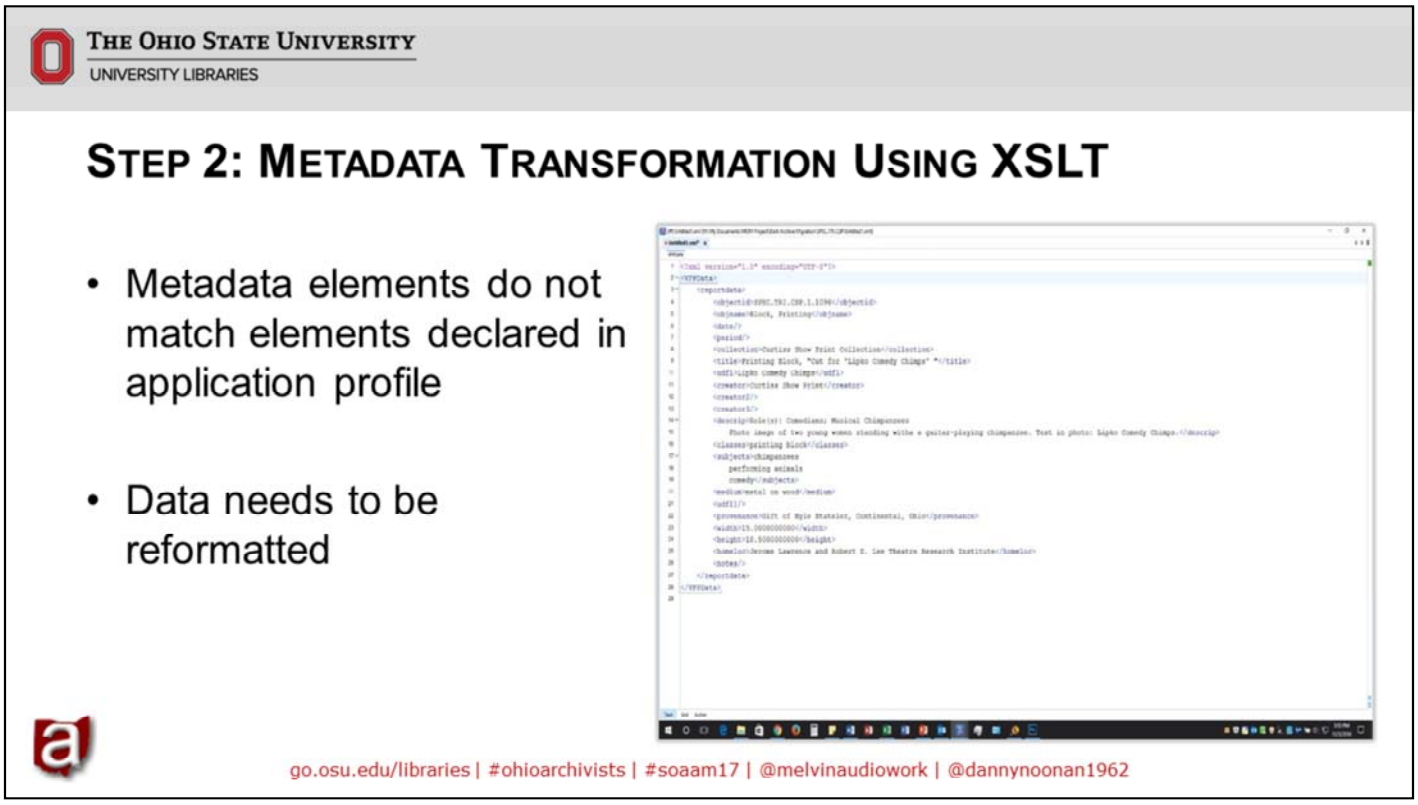
The first thing that needs to happen is to get our collection and item data out of the legacy system. Past Perfect has the ability to export a number of file formats. In our case, I need the data structured as XML.

For those of you who have not worked with Past Perfect before, there are four types of Collection modules in the system (Objects, Photos, Archives, and Library). Our special collections primarily utilizes the Objects and Archives Catalog.

The second thing that needs to happen prior to export, is determining what metadata fields are needed. For the collections in the Objects catalog, I exported all Object Identifiers, Object Names, Dates, Temporal data, Titles and Alternative Titles, Creator and Contributors fields, Descriptions, Classification, Subject Headings, Medium data, Language data, provenance data, measurements, location data, and any public or staff generated notes.

For the collections in the Archives catalog, I export the same fields previous mentioned with the addition of including Scope & Content Notes, and the Abstract Notes.

Next, it will be necessary to create an inventory spreadsheet which will include a column for the file path leading to the resources and a column for the corresponding Past Perfect Object Identifier.



Step 2, Metadata Transformation using XSLT.

So, after reviewing the exported data from Past Perfect, I noticed the data is well structured, but will need to be reformatted to comply with the Metadata Application Profile. This is were XSLT comes into play.

I created two transformations to deal these issue. Template one does five task. They:

1) Removes the Past Perfect Schema from the XML file;

2) Generates metadata elements and values for constant data for each item (i.e. resource type, collection identifiers);

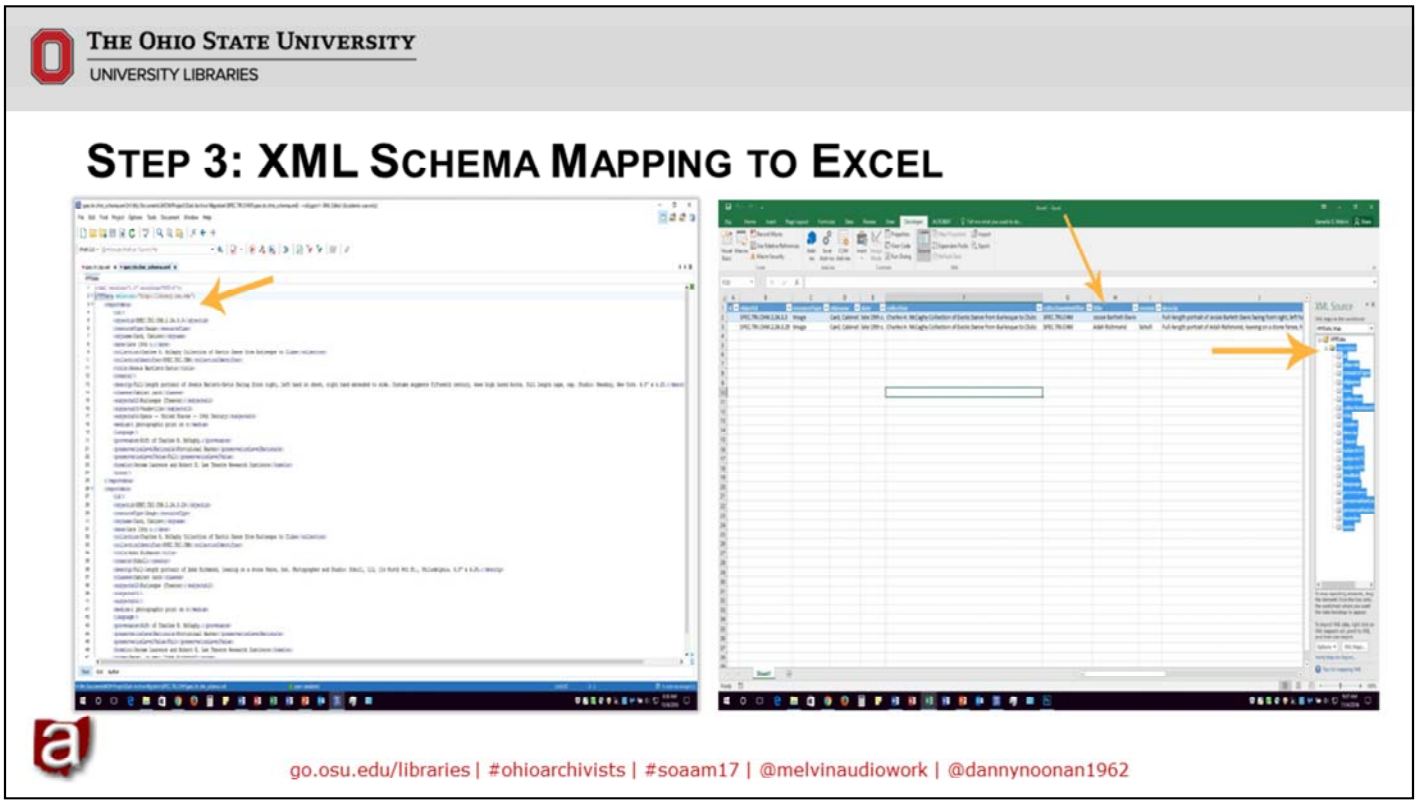
3) Rename all metadata elements (i.e. objname to VRA.workType; udf1 to DC.altTitle; udf11 to DC.language);

4) Parse out subjects and add them to their own element nodes (Past Perfect exports subjects heading as one long string); and

5) Normalize whitespace in the description field.

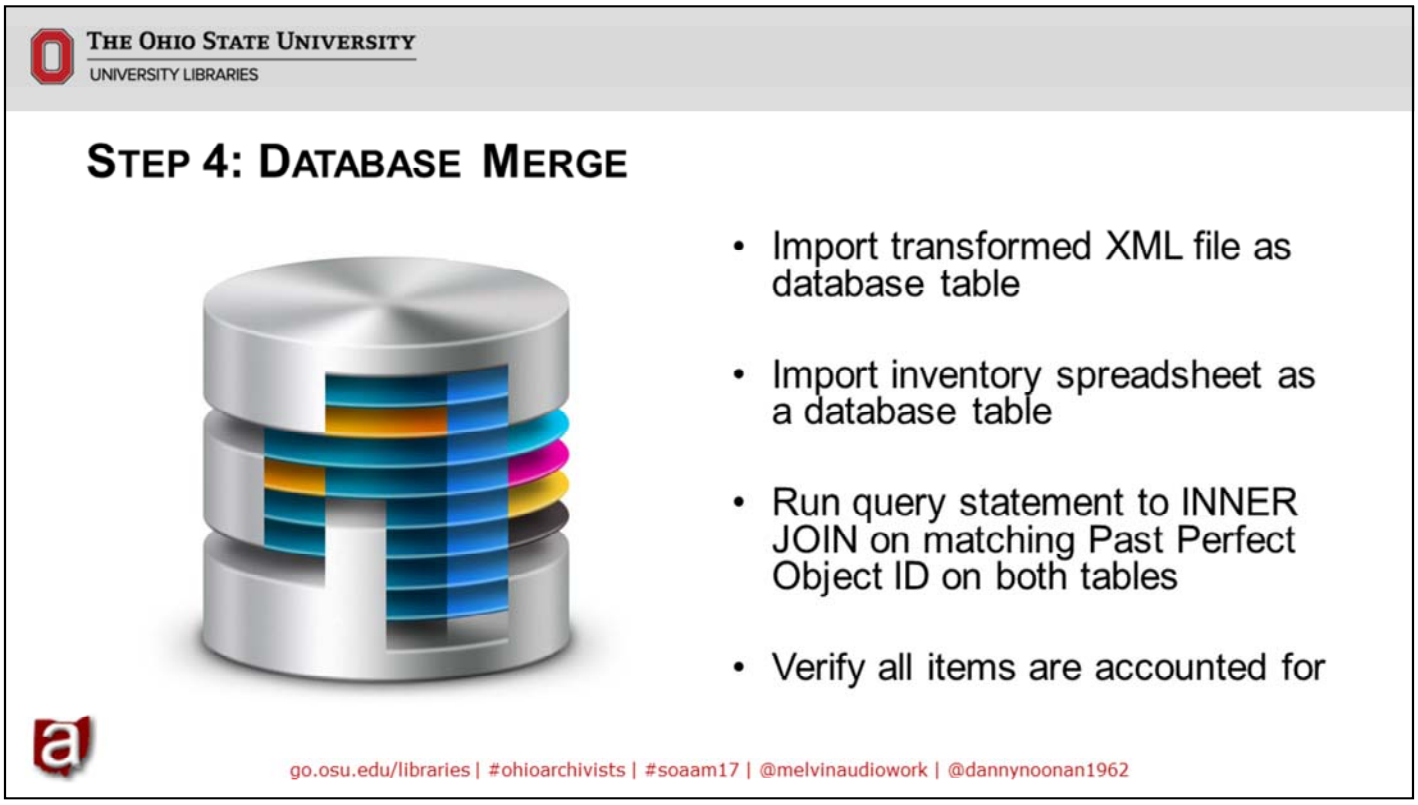
Template two:

Auto-numbers the subjects fields once the metadata is in their own nodes. This is needed to differentiate multiple subjects heading for an item once they are in a spreadsheet.



Step 3, XML Schema Mapping to Excel.

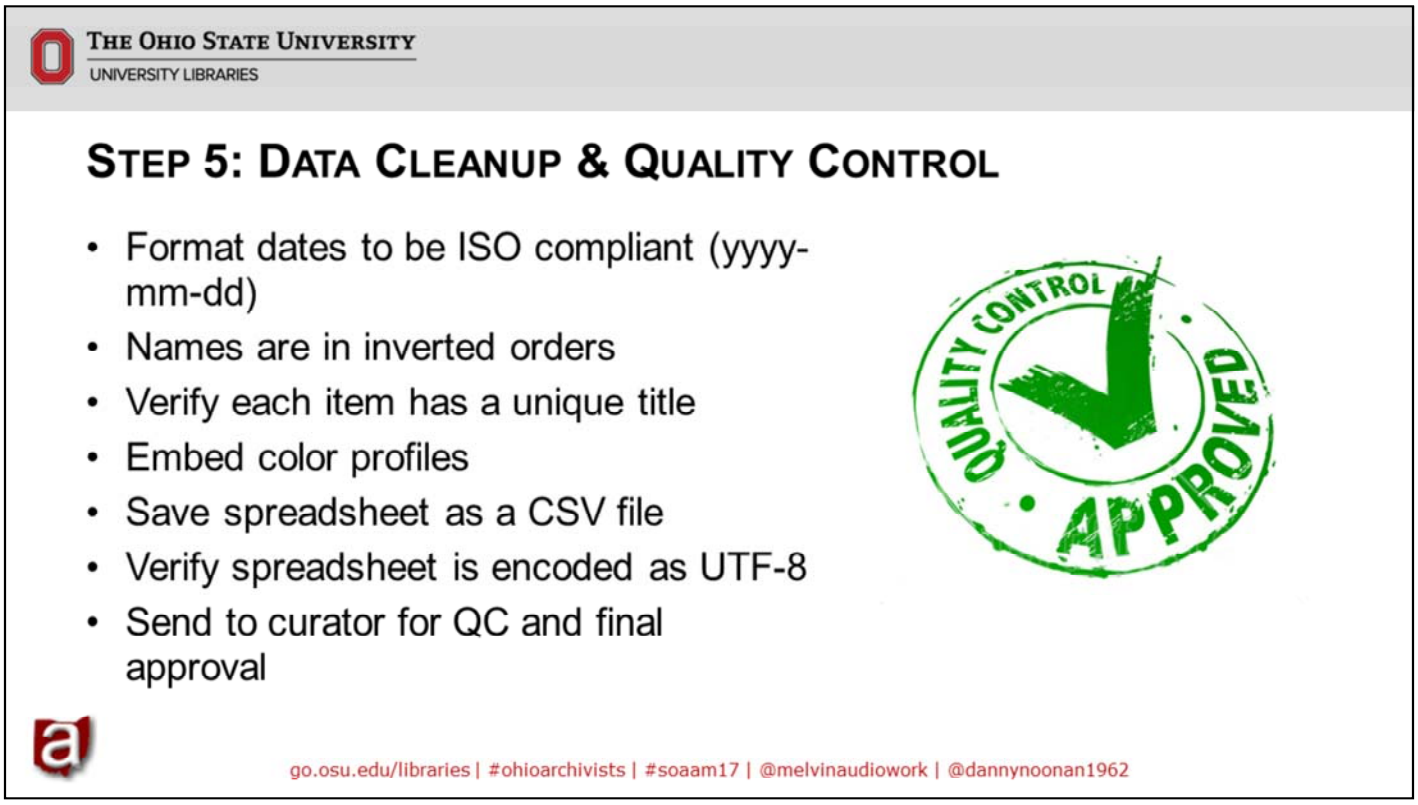
Before we can import a XML file of our collection and item data into a spreadsheet, Excel requires the XML structure to be mapped before import. I do this by taking two item records from the transformed XML file and save it as a new file. Once in Excel, we use it to map and the XML structure becomes the header of the spreadsheet. Once mapping is complete, you can then proceed to import your collection.



Step 4: Database Merge.

In this step, we import our two spreadsheets into a database and they become two tables. The first spreadsheet is the collection which includes item records. The second spreadsheet is our inventory file. Once both tables are in the database, we then run a query statement that INNER JOIN the two tables by matching on the Past Perfect Object Identifiers.

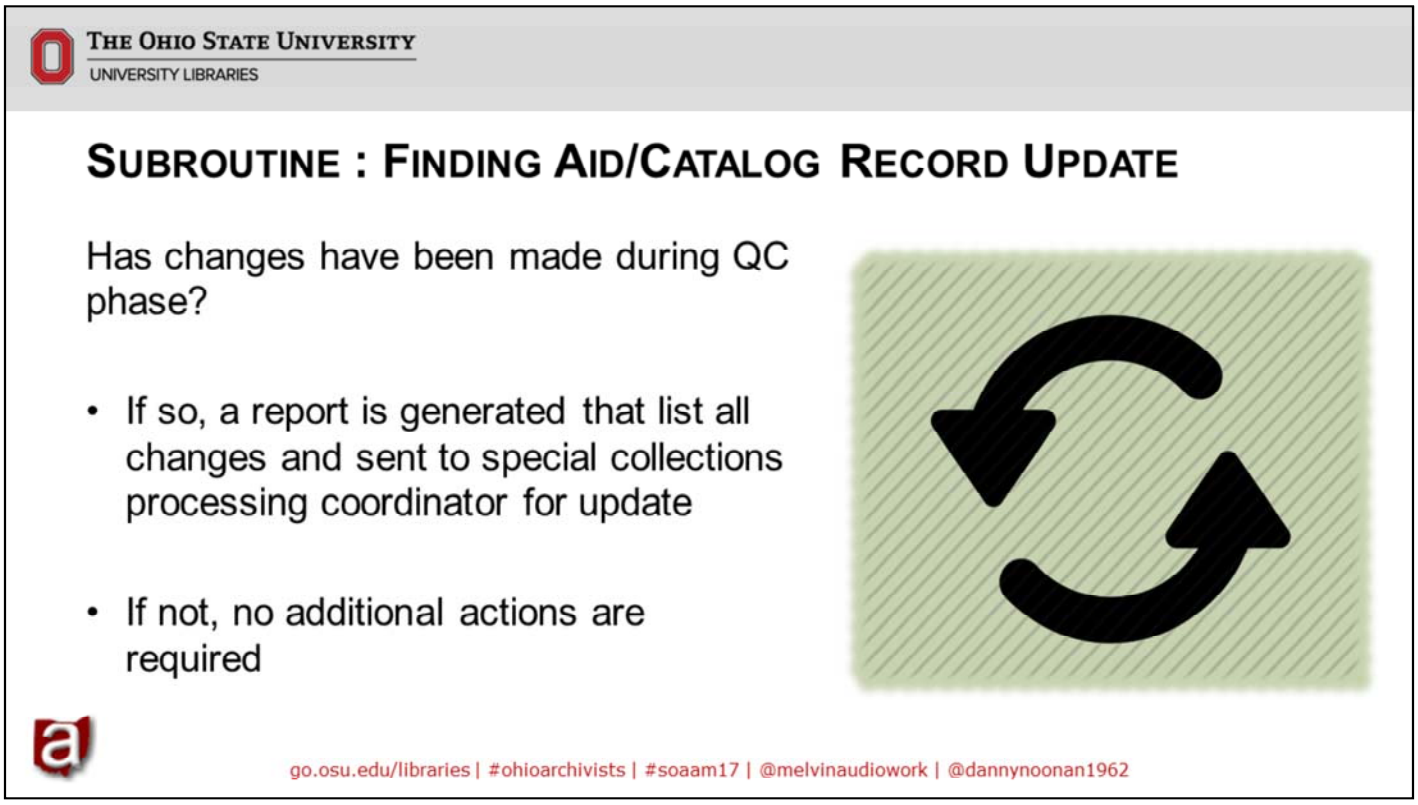
So after running the query, what we get is a new table that has the file-path merged with it corresponding metadata. This new table is what we will use for bulk ingest into the Master Objects Repository.



Step 5: Data Cleanup & Quality Control.

In this phase, we do the final cleanup on the data in the spreadsheet, which includes date formatting, name formatting, verifying all items has a unique title, and color profile mapping. Here, we need to save the spreadsheet as a CSV file and encode it as UTF-8.

Once the data has been cleaned and the file is saved, we send this final spreadsheet to the special collections curator to review the spreadsheet. Once they give us the ok, it time to ingest.

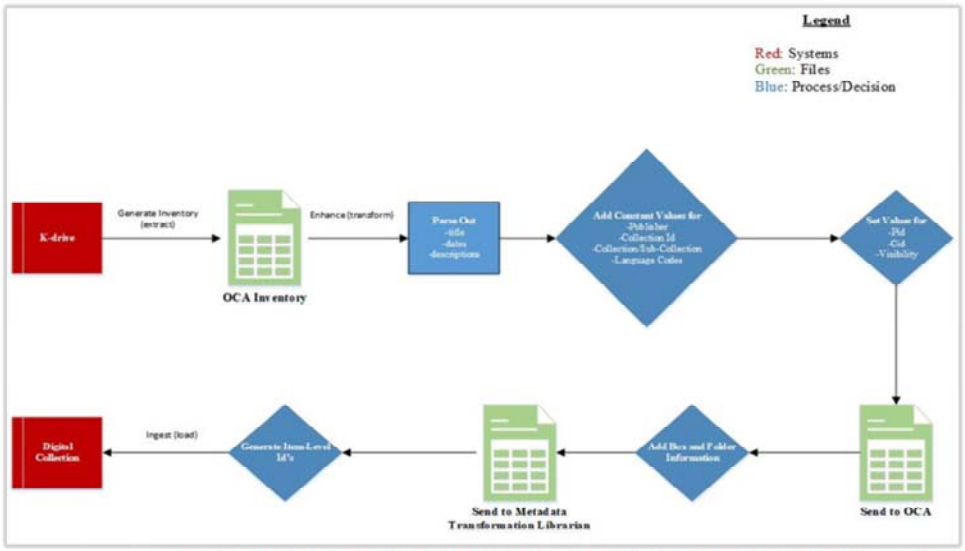


Subroutine, Update finding aid and or catalog record.

Now that QC has been completed, we want to make sure that if any changes to the content has been made, we want those changes to be reflected in the finding aid and/or catalog record if one exist. This task is done by forwarding a report to our special collections processing coordinator that list all changes and their staff will make all changes if needed.

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

MOR ARCHIVAL INVENTORY EXTRACT TRANSFORM LOAD WORKFLOW



go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

STEP 1: GENERATE INVENTORY

- Generate inventory
- dir /s/b/o:n > inventoryFile.xls

go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

Step 1, Generate inventory.

The first thing that needs to happen is to create an complete list of files in the directory. The script illustrated above is executed from the command line and will create a excel file listing the complete file path for all folders, sub-folders, and files in that location.

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

STEP 2: PARSE AND REPURPOSE METADATA

- **Trim File Name from File Path**

=TRIM(RIGHT(SUBSTITUTE(B2, "\",REPT(" ",100)),100))

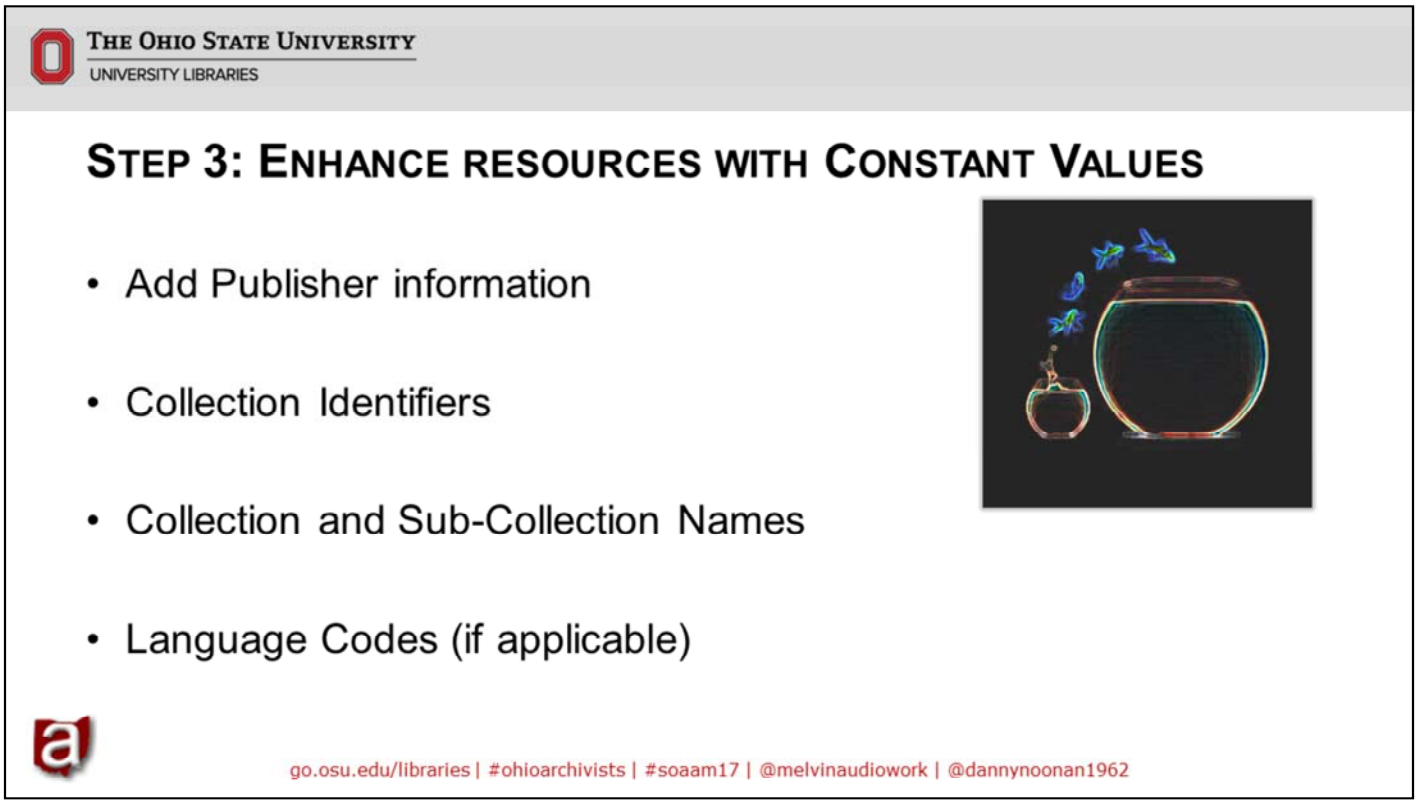
- **Extract Date from String**

=MID(B2,SEARCH("????-??-??",B2),10)

go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

Step 2, Parse and Repurpose Metadata.

Now that a complete inventory is in an Excel file, we can parse out useful descriptions and repurpose it. In this case, the file names are descriptions or titles and will need to be mapped to its corresponding metadata field. In addition to titles, the date created may also be in the file path.



Step 3, Enhance resources with Constant Values.

In this stage, we want to enhance the record with additional information. Here we will add publisher information, collection identifiers, collection and sub-collection name, and language codes if applicable.

In the case where we are ingesting complex objects, we need to assign the visibility status and add parent and child identifier to all associate items within a work.

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

STEP 4: ADD BOX AND FOLDER INFORMATION

- Add Box Numbers
- Add Folder Numbers

go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

Step 4: Add Box and Folder Information.

In step 4 we add box and folder information.

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

STEP 5: GENERATE ITEM-LEVEL IDENTIFIERS

Concatenate

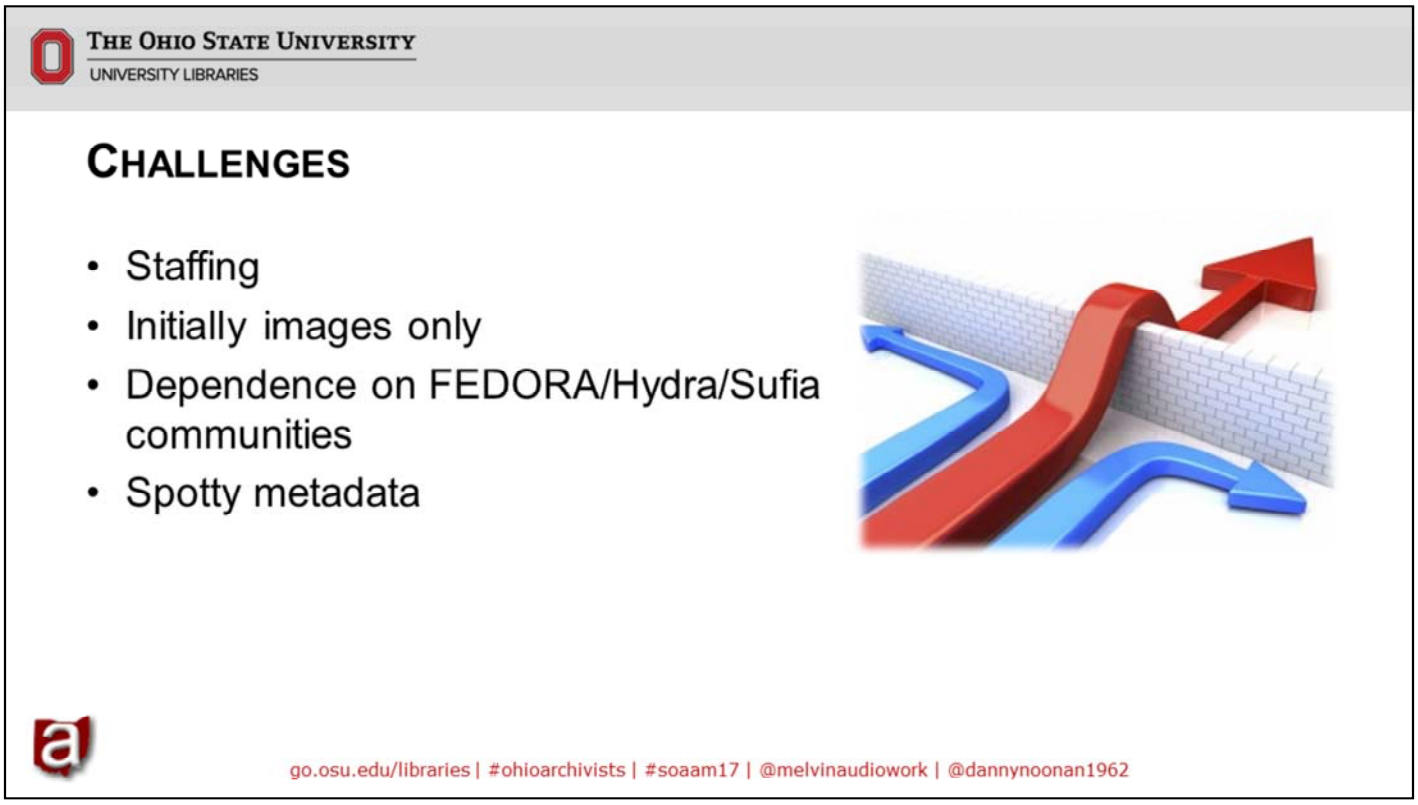
- Collection Name
- Box Number
- Folder Number
- File Name

"kilroy_box1_fdr36_2010-09-01"

go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

Step 5: Generate Item-Level Identifiers.

In step 5 we generate item-level identifiers by concatenating the collection name, with box and folder numbers and the file name.



In conclusion: I'd just like to point out a few of the challenges we have had/still have:

Staffing: Both for the development effort, as well as the curatorial and metadata efforts, we just haven't had enough staff to make it happen as fast as anyone would like to see it accomplished. Developer churn created issues early on, but we are relatively stable and staffed-up now. As for metadata transformation, our metadata transformation librarian is a term appointment and we will lose him within the next six months. And the archivists and curators try to find time to create metadata, while doing the rest of what they do—hence a need for clear prioritization.

Complicating the migration prioritizations was our self-imposed limitation for images only. That limitation recently has been lifted, but we are still just beginning to experiment with documents and complex objects. We will not have true ordered complex object capabilities until sometime this summer. Additionally, we have curator pushback that the system does not support enough archival hierarchy for appropriate contextual search and retrieval.

Another challenge has been a dependence on the FEDORA, Hydra and Sufia communities. This is not meant to throw slings and arrows at them, its just that without the expertise on staff we are reliant on community development for all the cool stuff we want to do—or sometimes basic functionality that our

archival/curatorial staff expect.

Lastly, the one thing that will keep us adrift is chasing down that elusive treasure, METADATA.

Thank you.

THE OHIO STATE UNIVERSITY
UNIVERSITY LIBRARIES

CONTACT INFO:

Darnelle Melvin

Metadata Transformation Librarian, Visiting Assistant Professor
Special Collections Description & Access
University Libraries
The Ohio State University
320H 18th Avenue Library
175 W. 18th Avenue
Columbus, OH 43210
melvin.372@osu.edu
614.247.4689
@melvinaudiowork

Dan Noonan

Digital Resources Archivist, Assistant Professor
University Archives
University Libraries
The Ohio State University
320B 18th Avenue Library
175 W. 18th Avenue
Columbus, OH 43210
noonan.37@osu.edu
614.247.2425
@dannynoonan1962

go.osu.edu/libraries | #ohioarchivists | #soaam17 | @melvinaudiowork | @dannynoonan1962

Thank you.