



THE OHIO STATE UNIVERSITY

UNIVERSITY LIBRARIES

Gray Digital Preservation Repository Workflow

[as of 2024.01.18]

0	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	0	1	0	1	0	1	0	0	1	0	1	0	0
0	1	1	1	1	1	1	0	0	1	0	0	1	0	1	1	0	1	1	1
1	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	0
1	0	1	0	1	1	0	0	0	1	1	0	1	0	1	1	0	0	1	1
0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0
0	1	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	1	1	1	0	0	0	0	0	1	1	1	1	1	1	0	1	1	1
1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
1	1	0	0	0	1	1	0	1	1	0	1	0	1	0	1	0	1	0	0
0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
1	0	0	0	1	0	0	0	1	0	1	0	1	1	1	1	0	0	0	0
0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1

This document is version of the workflow that has been redacted for information security purposes.

Contents

Introduction	4
Ingest Information Set Up.....	6
A Note About File Names	6
Ingest Documentation Folder	6
Remote Desktop.....	8
File Manifest, Characterization and Checksums	9
DROID.....	9
Shortcut.....	10
Launch DROID and set Preferences.....	12
Create Profile.....	14
Adding Content to Profile	14
DROID Results	16
DROID Report and Export.....	18
PII-Review	24
Optical Character Recognition (OCR)	24
Launch Quick PDF-OCR application	25
Select content to be OCR'd.....	26
Process Document.....	27
Results	29
Identifying PII Data with Bulk Extractor.....	32
Launch Bulk Extractor	32
Run bulk_extractor.....	32
Required Parameters	34
Scanners.....	36
Scanner Controls	37
Submit Run.....	38
Results	38

Bagging	41
Bagger/BagIt.....	41
Shortcut.....	42
Setup Bagger Profile	44
Run Bagger	45
Create a Bag	46
Adding a Payload	48
Adding Bag-Info.....	49
Saving Bag	50
Completed Bag	52
Ingest	55
Mapping █ bucket.....	55
Ingest	58
FederalID.....	60
Finding Aid Linkage.....	62
Archivist Toolkit.....	62
PastPerfect.....	62
Retrieval	62
Accessing the Gray Repo.....	63
Local Administrative Dashboard	66
Files Maintained Locally	67
Data Maintained in Gray Repo Admin Console	70
Admin Console Look and Feel.....	72
Entering Data into the Admin Console.....	74
What to do with files post-ingest	76
Resources	78

Introduction

Welcome to the explanation of and instructions to preparing and ingesting content into the Gray Digital Preservation Repository (Gray Repo). The Gray Repo is a "dim digital preservation archive" that provides no public access, and limited curatorial access to the University Libraries' digital objects stored within. This is in contrast to a "light archive" which provides public access, such as [Digital Collections](#), or a "dark archive" which only allows custodial access. The Gray Repo allows for curatorial deposit and retrieval, but no direct patron access. It is much more akin to a physical archival storage facility, much like our Book Depository, where items are stored on shelves in a regulated and well managed manner, appropriately described in conformance with accepted standards, however the public, and unvetted personnel are not allowed to wander the stacks. The Gray Repo is built utilizing [Fedora](#), an open source digital preservation solution, that has been installed in an Amazon Web Services (AWS) platform.

There are six (6) basic steps for preparing and ingesting the content that are also illustrated in Figure 1:

- File Manifest, Characterization and Checksums
- PII Identification
- Bagging
- Ingest
- Finding Aid Linkage
- Local "Administrative Dashboard"

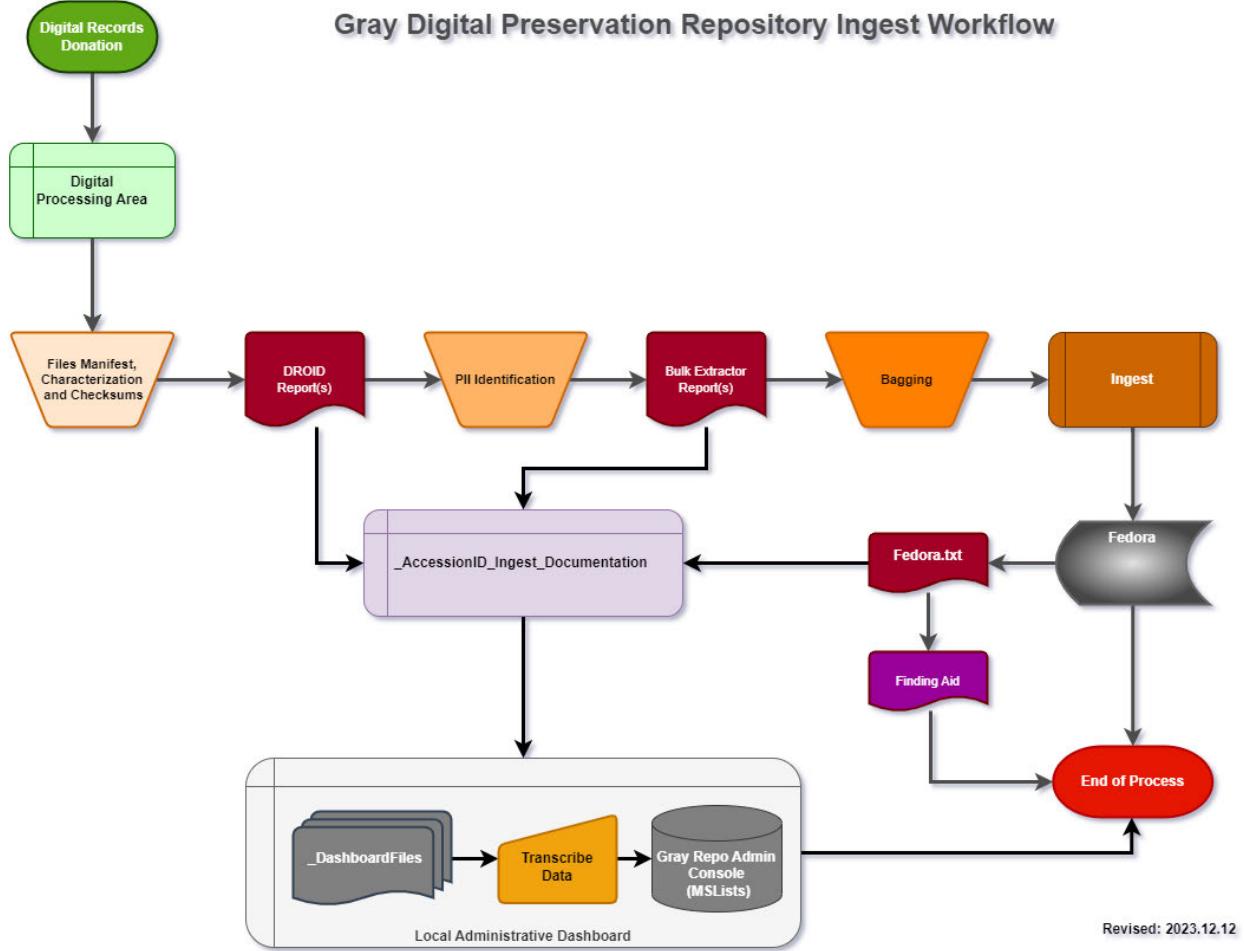


Figure 1: Gray Digital Preservation Repository Ingest Workflow Diagram

Ingest Information Set Up

A Note About File Names

For the purposes of this overall ingest workflow we need to have file names that are either camel-case or uses an underscore (_) as a spacer; the use of periods or dots (.) can produce unwanted results with some of the tools. Additionally, for reports that you generate and save, you will want include the accession ID. All files that are generated and to be saved during the ingest workflow should be prefixed with the accession ID. Below are some examples:

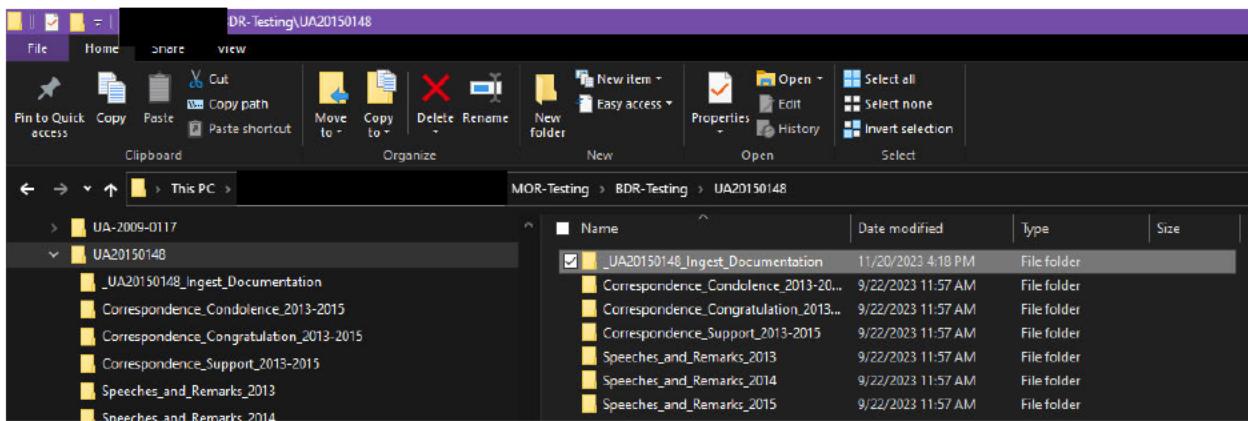
- CamelCase.xxx
 - UA20150148DROIDReport.xlsx
- Under_score.xxx
 - UA20150148_droid_report.xlsx
 - UA_2015_0148_droid_report.xlsx
- Camel_Case_Underscore.xxx
 - UA20150148_DROID_Report.xlsx
 - UA_2015_0148_DROID_Report.xlsx

While there are options, choose a method and be consistent.

Ingest Documentation Folder

Throughout this workflow we will be generating documentation about the folders and files that will be ingested into the Gray Repo. We need to create a new folder within the root folder of every accession that you will process to collect this information. To that end create a folder that follows this pattern:

- “_AccessionID_Ingest_Documentation”
 - _UA20150148_Ingest_Documentation
 - _UA_2015_0148_Ingest_Documentation

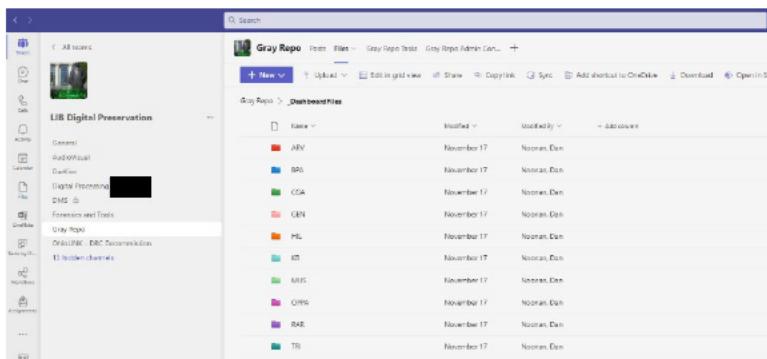


Ingest Documentation Figure 1: Example of Ingest Documentation folder naming

As can be seen in *Ingest Documentation Figure 1*, utilizing the “_” at the beginning of the folder name, allows it to rise to the top of the folder listing for easy recognition and access. This is the folder where we will collect the output of DROID, PII review process, and the [FederalID].txt. Files and the naming schema to be maintained here include:

- AccessionID.droid
- AccessionID_debug_file.txt (if necessary)
- AccessionID_DROID_Report.pdf
- AccessionID_DROID_Report.xlsx or AccessionID_DROID_Report.csv
- AccessionID_FederalID.txt (where FederalID = the actual FederalID)
- AccessionID_pii.txt
- AccessionID_ReadMe_YYYYMMDD.txt: This file should be created as necessary. We strongly encourage its inclusion to document issues encountered and/or actions taken throughout the process. You can create this with any text editor such as Notepad, Notepad++ or WordPad.

Ultimately, these files will be copied to the appropriate folder within the Gray Repo's [Local Administrative Dashboard](#) within the [LIB Digital Preservation>Gray Repo Team](#).

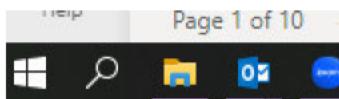


Ingest Documentation Figure 2: Gray Repo's Local Administrative Dashboard file storage

Remote Desktop

The University Libraries has setup a dedicated workstation with the necessary tools for conducting the content analyses and ingest into the Gray Repo. This workstation (██████████) can be accessed via Remote Desktop. This dedicated system has DROID installed.

To access ██████████ via Remote Desktop, click on magnifying glass in the Windows Task Bar.



Remote Desktop Figure 1: Windows Task Bar

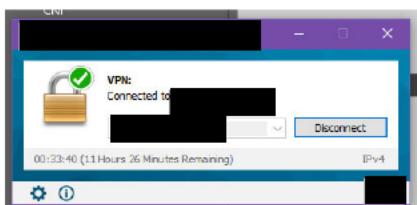
Then type *Remote* into the search bar – Remote Desktop Connection App should appear at the top of the list.

- You can click on the app to start immediately, or
- You can right click on the app and choose to “Pin to Start” and/or “Pin to Taskbar”; this will allow easier access in the future



Remote Desktop Figure 2: Remote Desktop interface

- You will be asked to use your name.# and pw credentials to login.
- If you are working remotely, you will have to be VPN'd into the ██████████ network using ██████████



Remote Desktop Figure 3: ██████████ interface

File Manifest, Characterization and Checksums

The first step in analyzing and processing our digital content for ingest into the Gray Digital Preservation Repository is to create a file manifest or list, allowing us to quantify the accession, determine what file formats we have and generate checksums for each file. We will utilize a digital forensics tool, [DROID](#), developed by The National Archives for the United Kingdom.

The file manifest provides detail in regards to how many folders and files we have, along with allowing us to quantify the size of the accession and ingest. A listing of the folders provides potential insight into the context of the records, and could be used to augment the finding aid. With the appropriate settings selected, DROID can examine and list the content in zipped file containers.

DROID generates a checksum for each file, essentially a digital fingerprint that can be utilized for future authentication, as well as determining if we have duplicate files. Should a file set indicate a significant number of duplicates, contact the Digital Preservation Department for assistance in de-duplicating the file set; we have additional tools that can assist in this process.

Further, DROID's core purpose is to conduct a file format characterization analysis, identifying not only file formats and which version they are, but can also identify file extension/file type mismatches. It provides linkage to [PRONOM](#), an on-line database about file formats and their supporting software, for deeper analysis if warranted.

Finally, as noted in the previous section, the output from the DROID analysis will be kept within the [Local Administrative Dashboard](#), and can be shared with patrons/researchers as a first look at the records prior to them being retrieved from the Gray Repo.

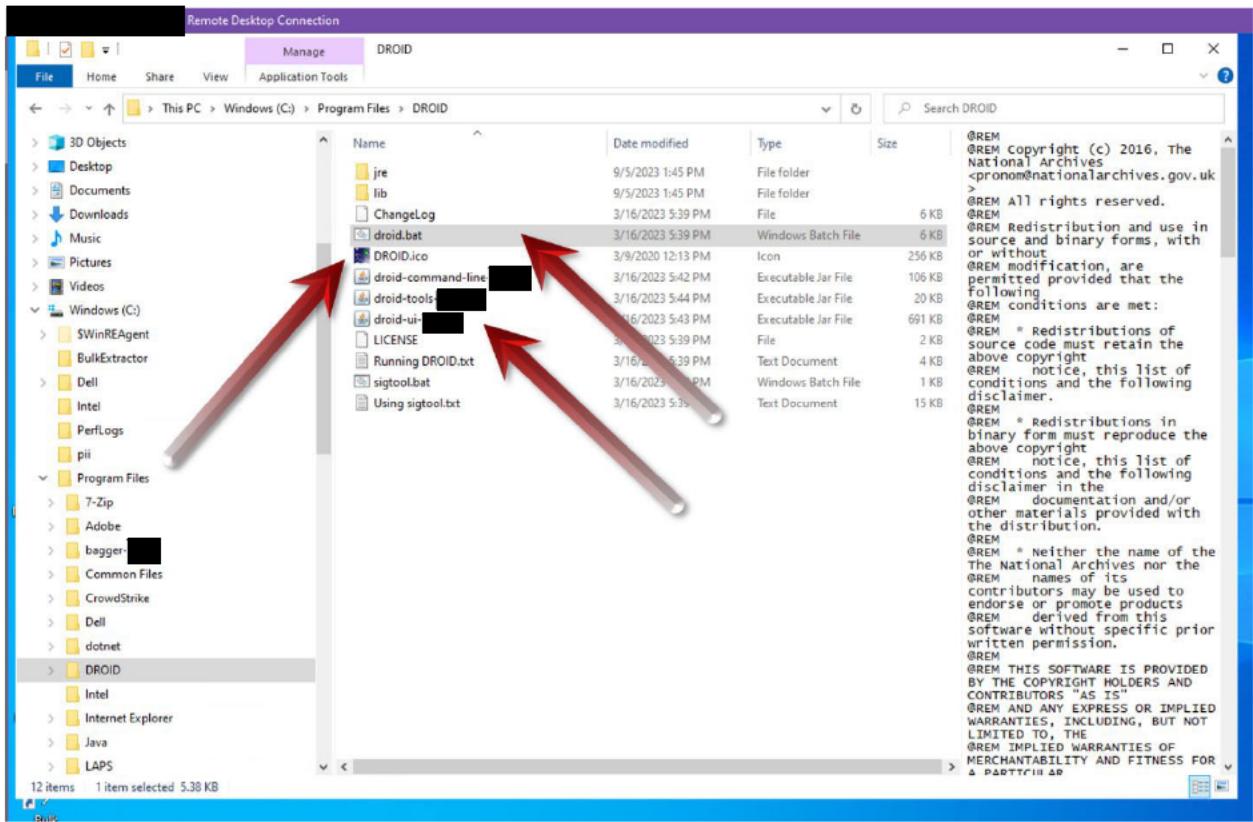
DROID



DROID (*Digital Record Object IDentification*) as previously noted is a file format identification tool developed by The National Archives UK. It additionally allows us to create a files manifest and generates checksums for fixity and de-duplication purposes. DROID has been installed on the [Remote Desktop](#) at C:\Program Files\DR0ID. When you use it for the first time, you will want to set up a desktop shortcut.

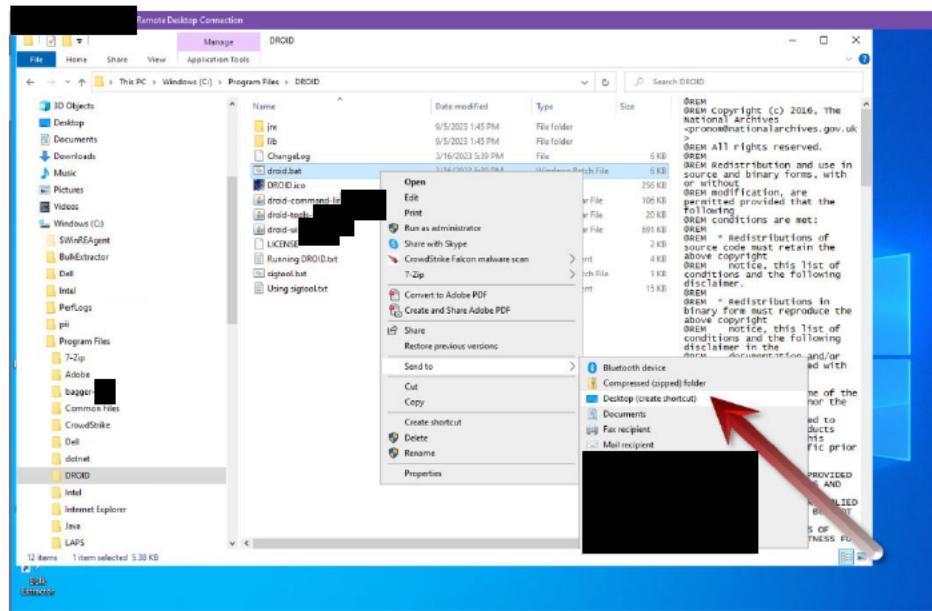
Shortcut

- Navigate to C:\Program Files\DROID
- While you can just “Double Click” on either the droid.bat or droid-ui-[REDACTED].jar (or more current version when installed) files to launch the program, it will be easier in the long run to have a desktop shortcut



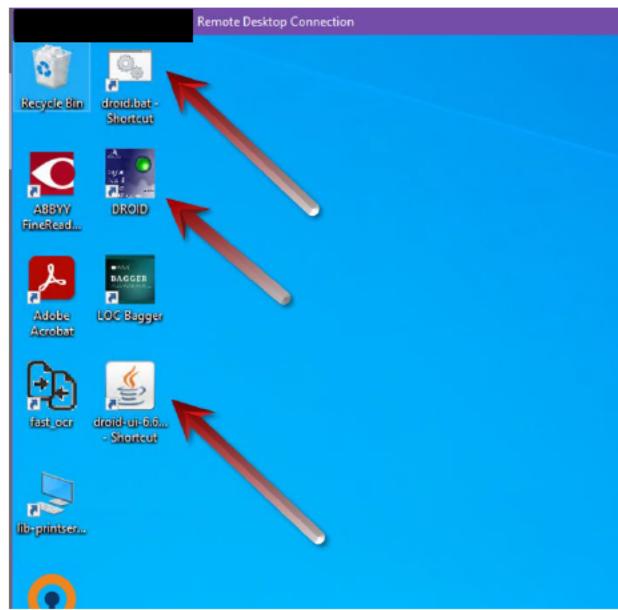
DROID Figure 1: Executable files and icon

- “Right Click” on either the droid.bat or droid-ui-[REDACTED].jar files
 - Click on “Send to”
 - Click on “Desktop (create shortcut)”



DROID Figure 2: Create desktop shortcut

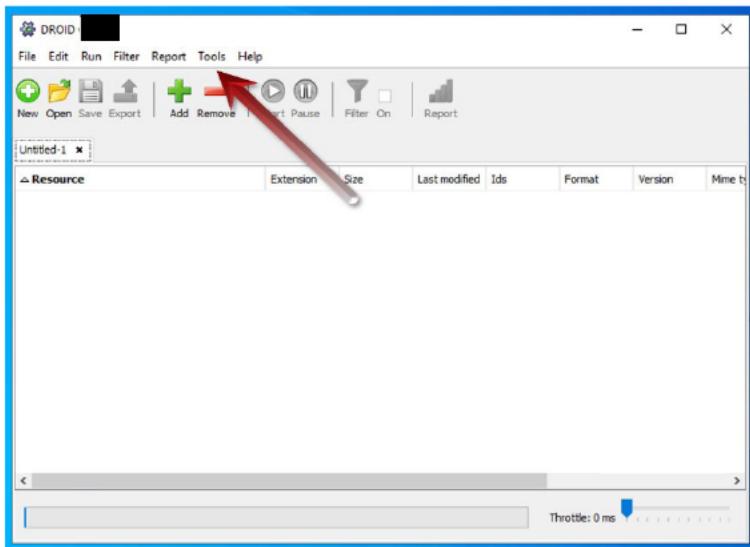
- Depending upon which file you chose, a desktop shortcut of either “droid.bat – Shortcut” or “droid-ui-[REDACTED] – Shortcut” will be created.
- You can rename the shortcut to just “DROID” and change its icon; contact the Digital Preservation Department for assistance.



DROID Figure 3: Desktop shortcut examples

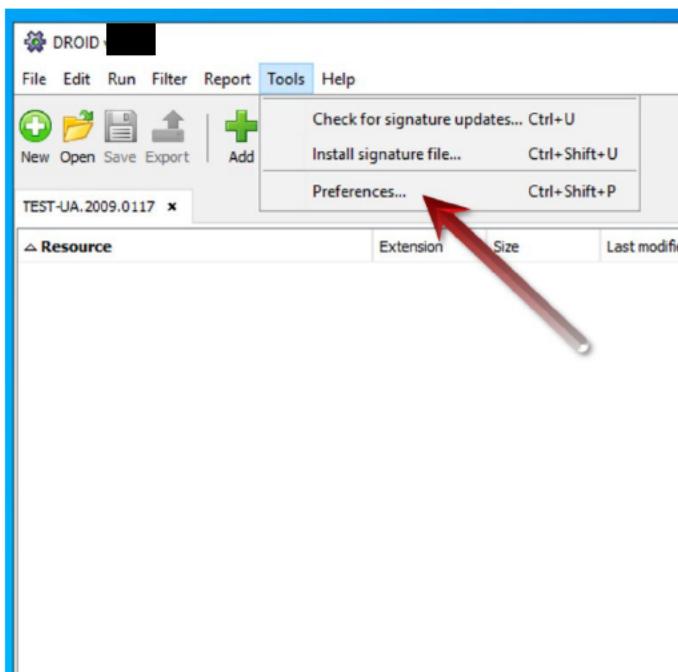
Launch DROID and set Preferences

- Launch DROID either by double clicking the desktop shortcut or navigating to C:\Program Files\DRDID and double clicking on either the droid.bat or droid-ui-[].jar
- Click on Tools in the navigation bar



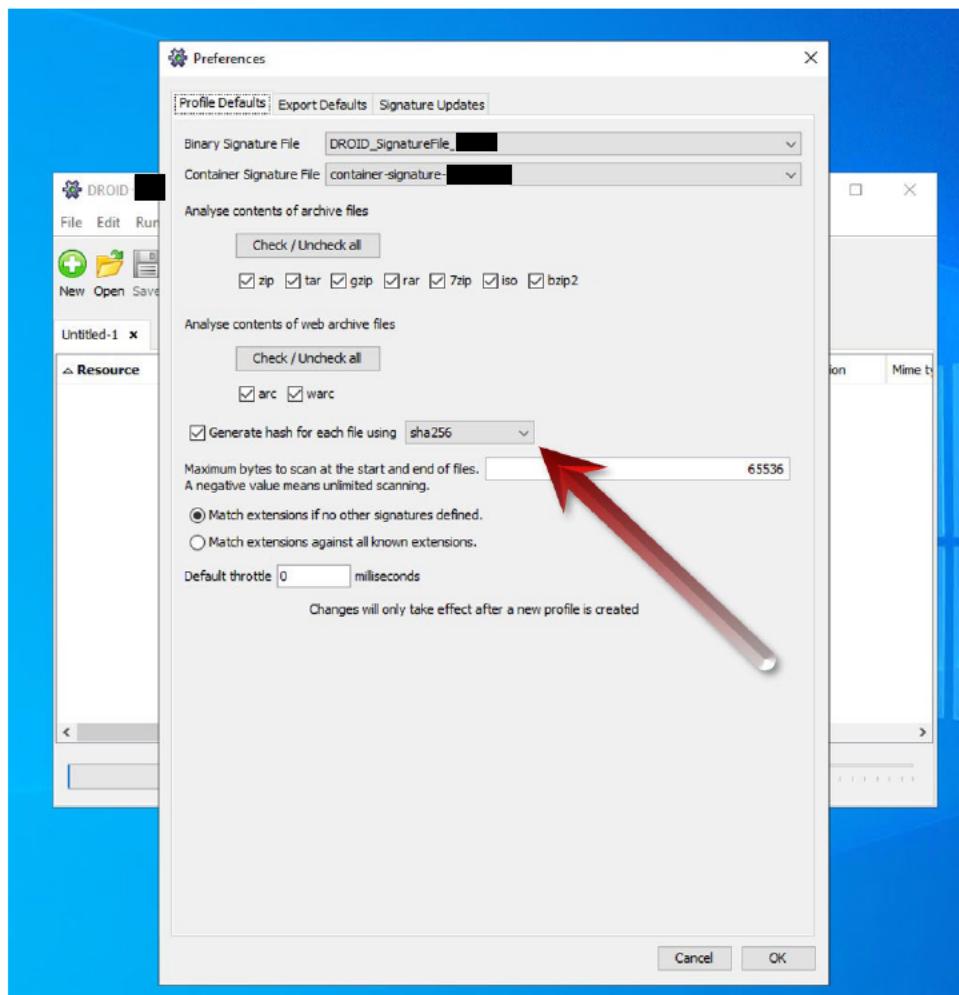
DROID Figure 4: Select Tools

- Click on Preferences



DROID Figure 5: Select Preferences

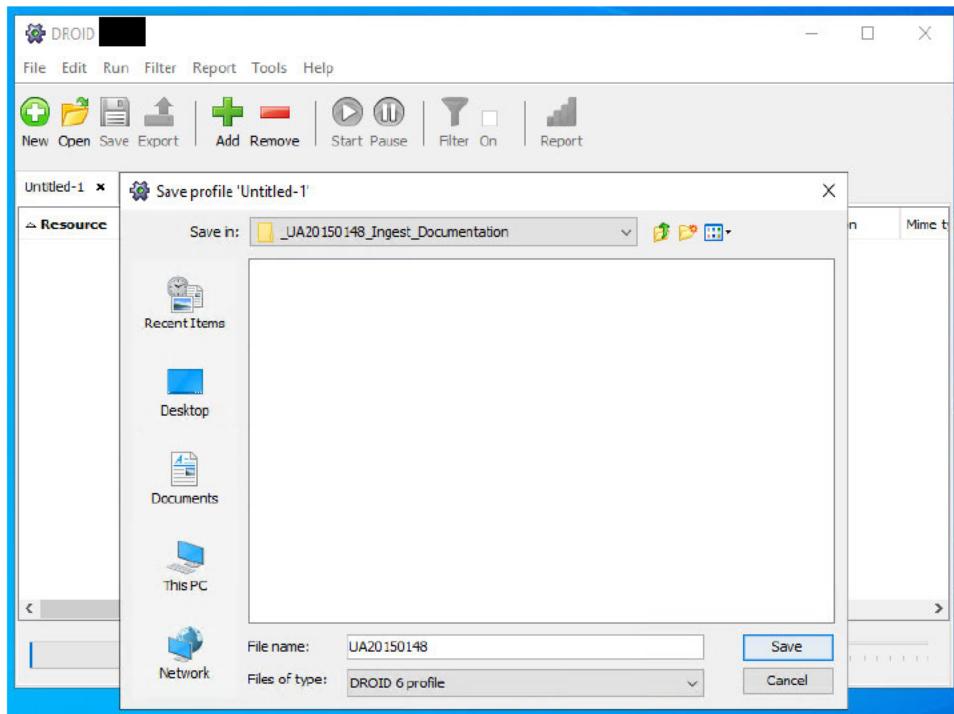
- Review the preferences in the Profiles Default tab
 - Make sure zip, tar, gzip, rar, 7zip, iso and bzip2 are all checked, as this allows DROID to examine the contents within those containers
 - Similarly make sure arc and warc are checked as this allows these web archiving containers to be scanned.
 - Most importantly make sure “Generate hash for each file using” is checked. From the pull-down menu select “sha256” if it is not already chosen.
 - Click OK
 - If the “Generate hash...” box had not been checked, after clicking OK you will need to exit DROID and restart it, or it will not remember the preference changes.



DROID Figure 6: Choosing preferences

Create Profile

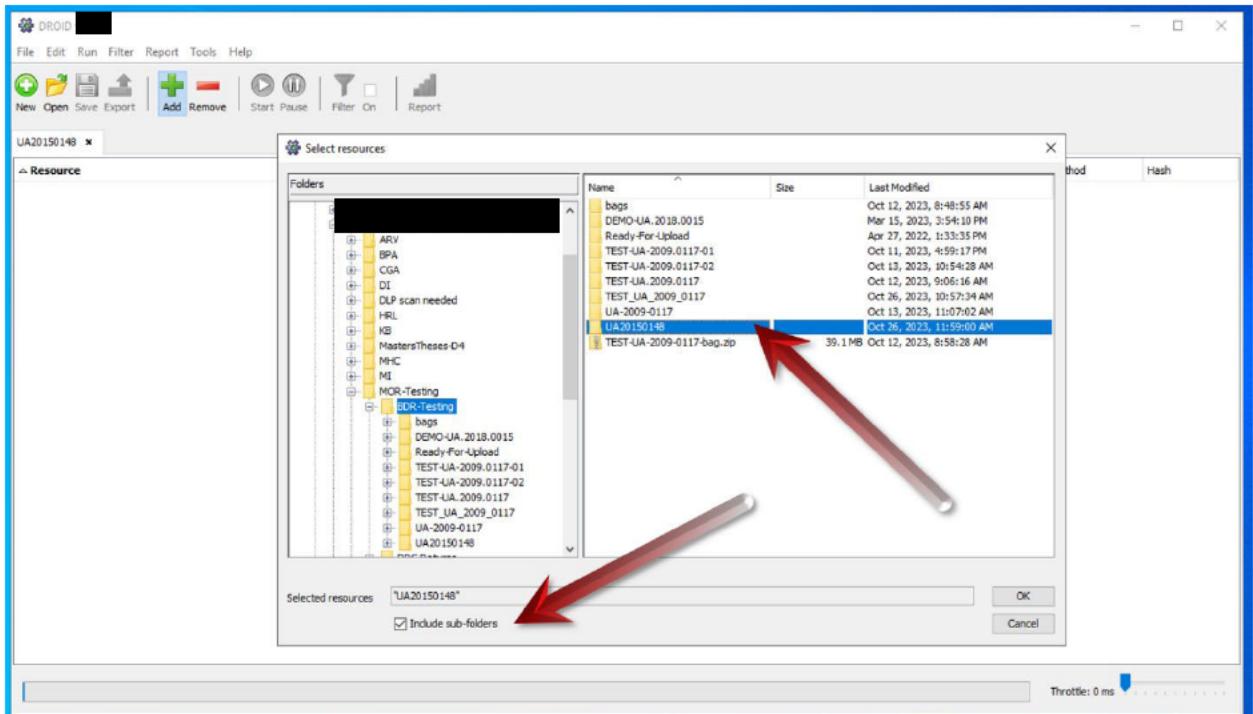
- After verifying preference and rebooting DROID if necessary you will need to create the profile for the current ingest
- When DROID boots, it creates an “Untitled-1” profile; you will want to save this profile, renaming it in the process.
- Click on File and then Save As. Navigate to the “_accessionID_Ingest_Documentation” folder you created in the root folder for the accession you will be analyzing. Save the profile within accordance with the file naming schema outlined above.



DROID Figure 7: Saving DROID profile

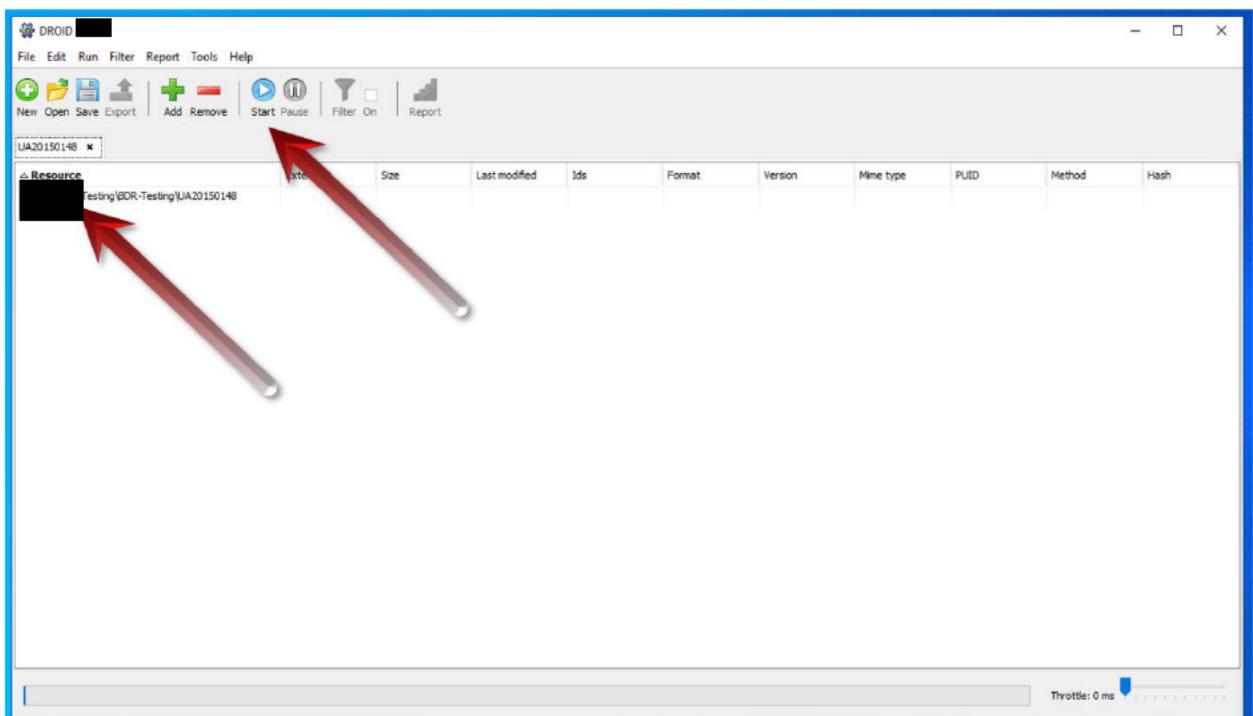
Adding Content to Profile

- Once the profile has been saved, we need to add content that will be analyzed.
- Click on the green plus-sign.
- This brings up a file navigation panel.
- Navigate to the appropriate folder to be analyzed; select it.
- Make sure the “Include sub-folder” is checked.
- Click OK



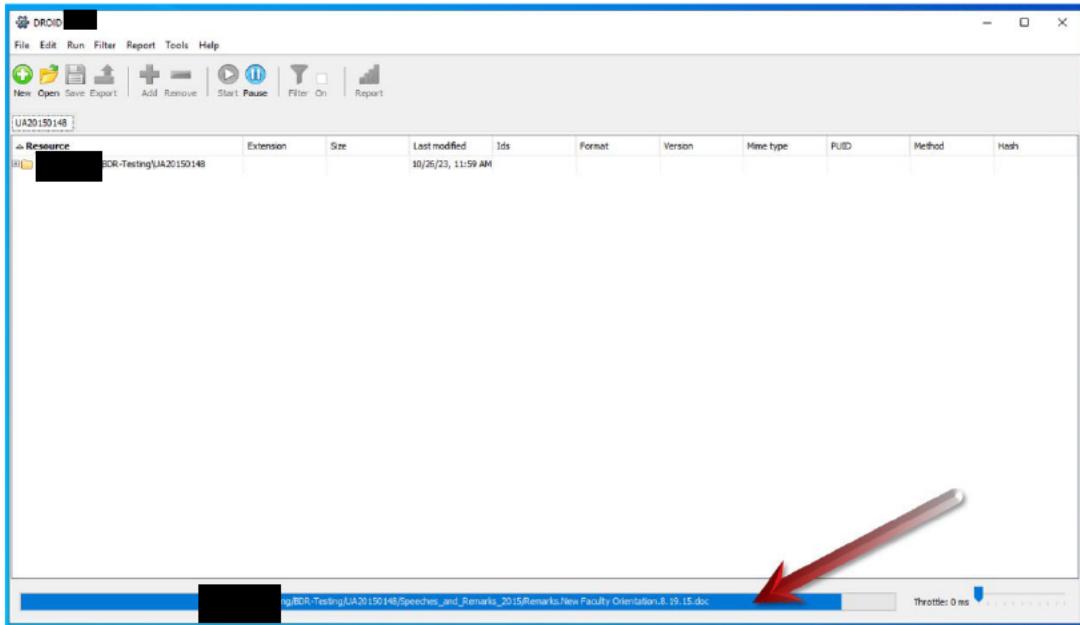
DROID Figure 8: Selecting folder(s) to be analyzed

- The file path should now show up underneath Resource
- Click the Start button to kick off the DROID analysis



DROID Figure 9: Start DROID analysis and progress bar

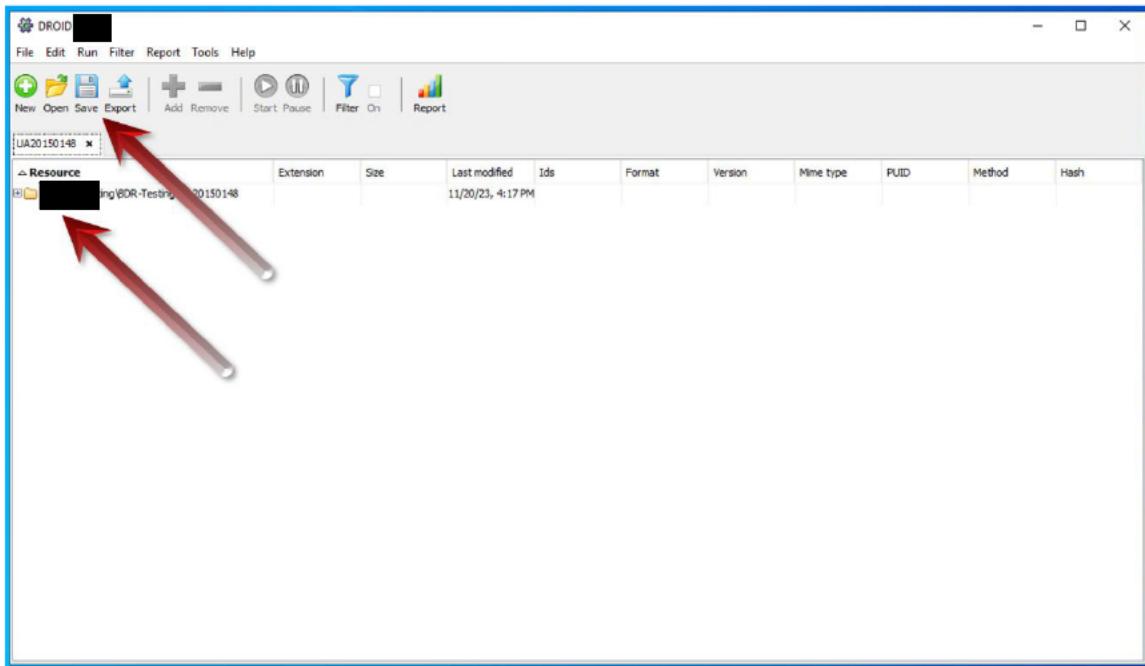
- A progress bar at the bottom indicates the status of the DROID tool's review



DROID Figure 10: DROID Progress Bar

DROID Results

- Once DROID has completed its analysis, you will notice that the file path folder has changed from white to manila yellow
- At this point Save the profile again



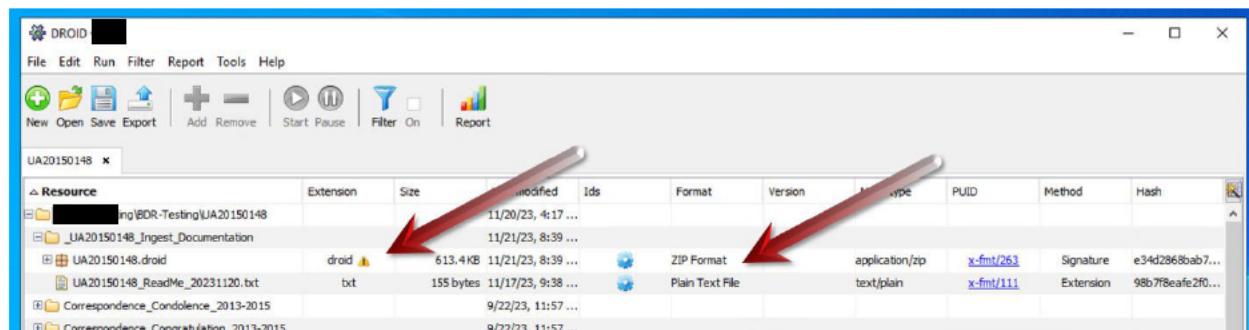
DROID Figure 11: Initial DROID results view

- By clicking on the plus sign next to the folder, you can open the folder similar to Windows Explorer. You can further navigate and explore the results by clicking open additional folders.

DROID Figure 12: DROID results details at the folder level

DROID Figure 13: DROID results details at the file level

- In DROID Figures 11 and 12 we can see examples of the details of the analysis that include:
 - File name
 - File extension
 - Size
 - Last Modified Date
 - Format and Version
 - Mime Type
 - PUID (PRONOM Unique Identifier) – a hyperlinked code to the PRONOM database
 - Hash (checksum)
- The DROID profile could be used to do some initial arrangement, description and scooping research of the files by reviewing the folder structure.
- It can also be used to determine if there are potential for file format anomalies, such as exotic file format types that are not recognized in the PRONOM database, which may present preservation issues; or file extension and format mismatch errors. As can be seen in DROID Figure 13, the file format extension “droid” is followed by a yellow triangle with an “!” in it. This indicates the extension does not match the format, which in this case is indicated as a ZIP Format. In this instance it is not an issue, but maybe you will have files that have a .tif or .pdf extension that are actually .jpg and .doc respectively. It would be problematic to render these if not fixed.



The screenshot shows the DROID application window. At the top is a toolbar with icons for New, Open, Save, Export, Add, Remove, Start, Pause, Filter On, and Report. Below the toolbar is a navigation bar with the path 'UA20150148 x'. The main area contains a table with columns: Resource, Extension, Size, Modified, Ids, Format, Version, Mime Type, PUID, Method, and Hash. A red arrow points to the 'Extension' column of the file 'UA20150148.droid', which is listed as 'droid' with a yellow warning icon. Another red arrow points to the 'Format' column, which is listed as 'ZIP Format'. The table data is as follows:

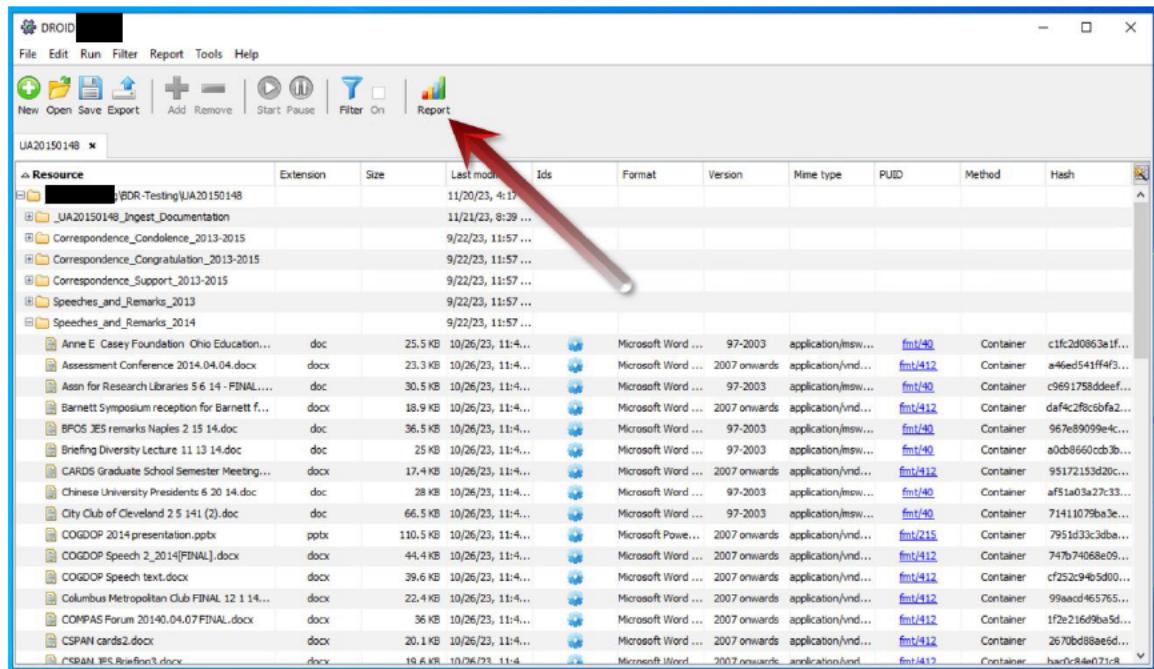
Resource	Extension	Size	Modified	Ids	Format	Version	Mime Type	PUID	Method	Hash
UA20150148			11/20/23, 4:17 ...							
_UA20150148_Ingest_Documentation			11/21/23, 8:39 ...							
UA20150148.droid	droid !	613.4 KB	11/21/23, 8:39 ...		ZIP Format		application/zip	x-fmt/263	Signature	e34d2868ab7...
UA20150148_ReadMe_20231120.txt	txt	155 bytes	11/17/23, 9:38 ...		Plain Text File		text/plain	x-fmt/111	Extension	98b7f8afe2f0...
Correspondence_Condolence_2013-2015			9/22/23, 11:57 ...							
Correspondence_Congratulation_2013-2015			9/22/23, 11:57 ...							

DROID Figure 14: File extension and format mismatch example

DROID Report and Export

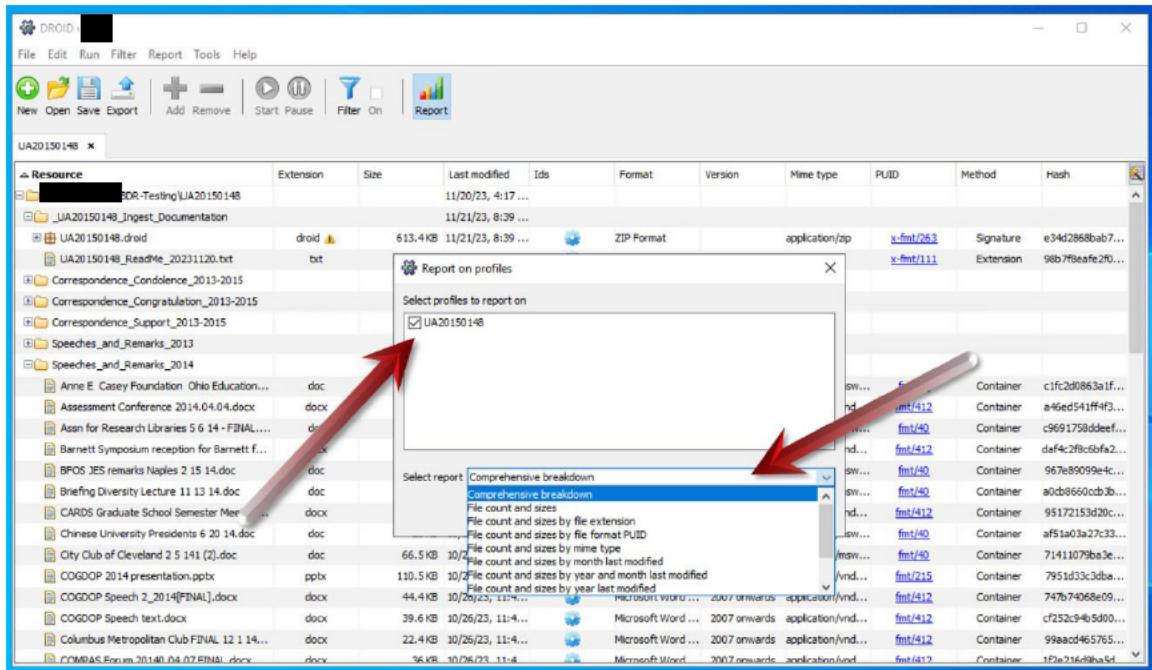
- To conduct deeper statistical and deduplication analyses, you have two options to choose from:

- Generate a report (required) – The report functionality can do a good job of aggregating and totaling files in various categories including format, mime types and file sizes, but is visually “clunky”.
- Export the data to a csv (required) – The csv or comma-separated values file, can be opened in Microsoft Excel or another spreadsheet or database program. As such the data can be filtered, sorted, totaled and formatted for additional analyses. It is especially useful for comparing hash values to determine if there are file duplication issues.
- To generate a report, click on the Report icon



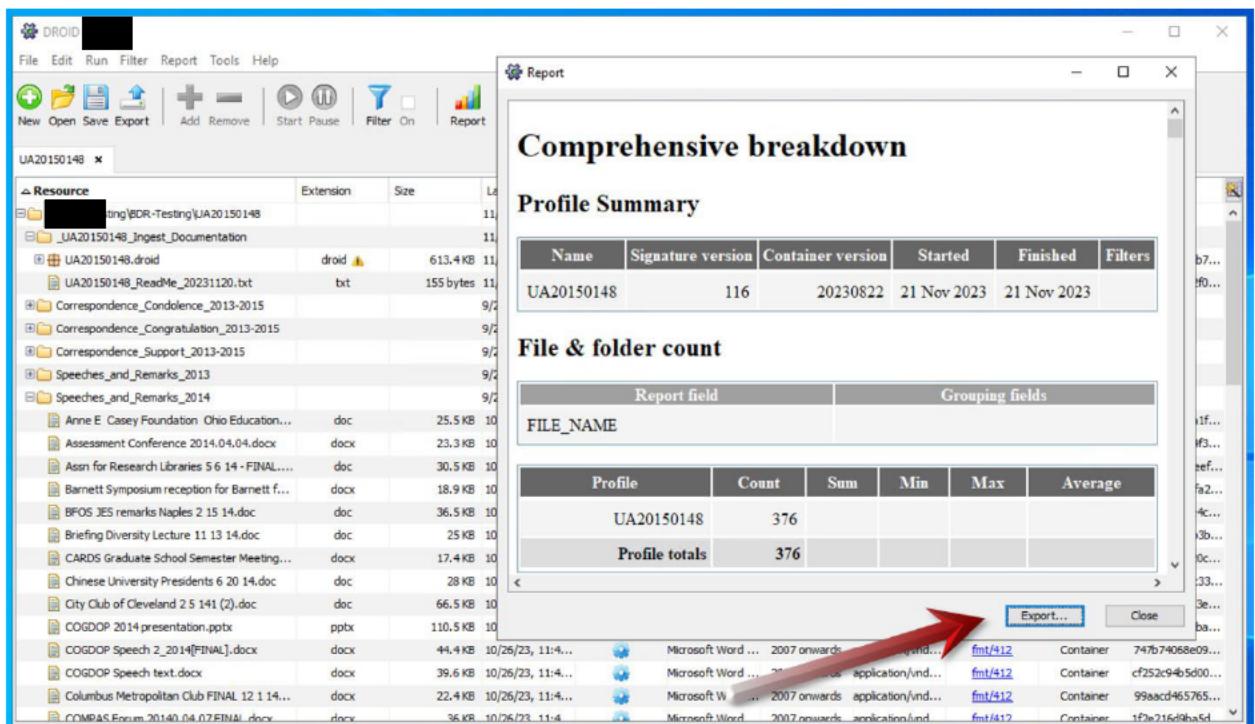
DROID Figure 15a: Select Report

- This will open a dialog window; select your profile. Additionally, you have report options to choose from; in the example we have chosen the comprehensive breakdown report.

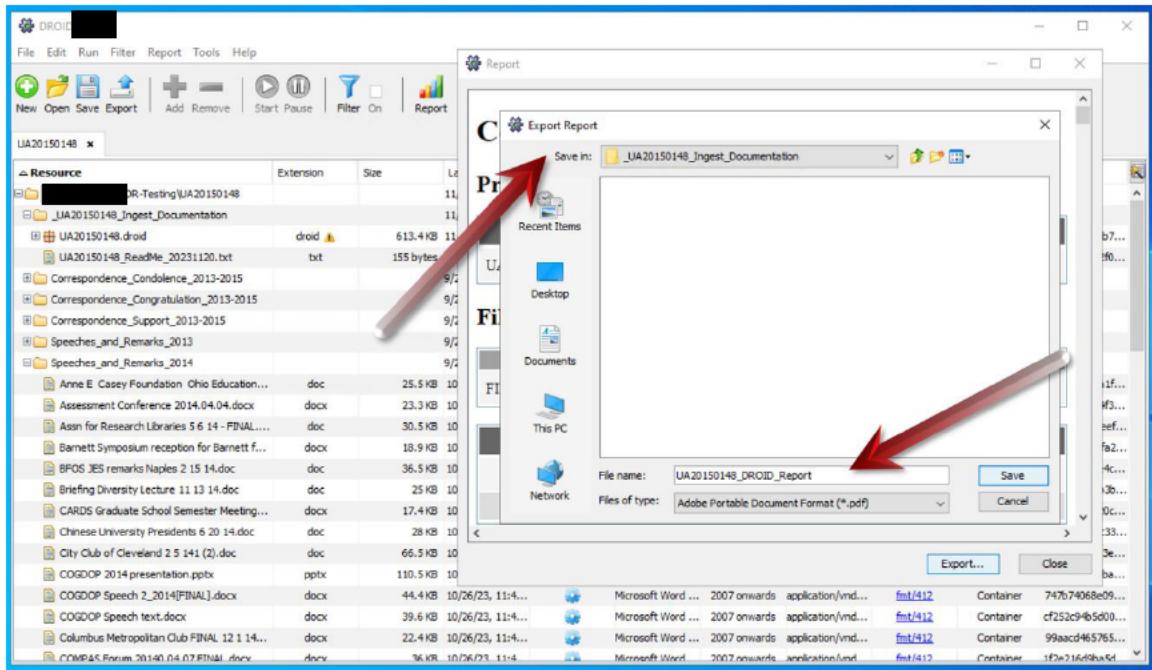


DROID Figure 165b: Profile and report type selection

- The resultant report can then be exported as a PDF. Save this report in the “_accessionID_Ingest_Documentation” folder. Close the dialog window after export is complete.

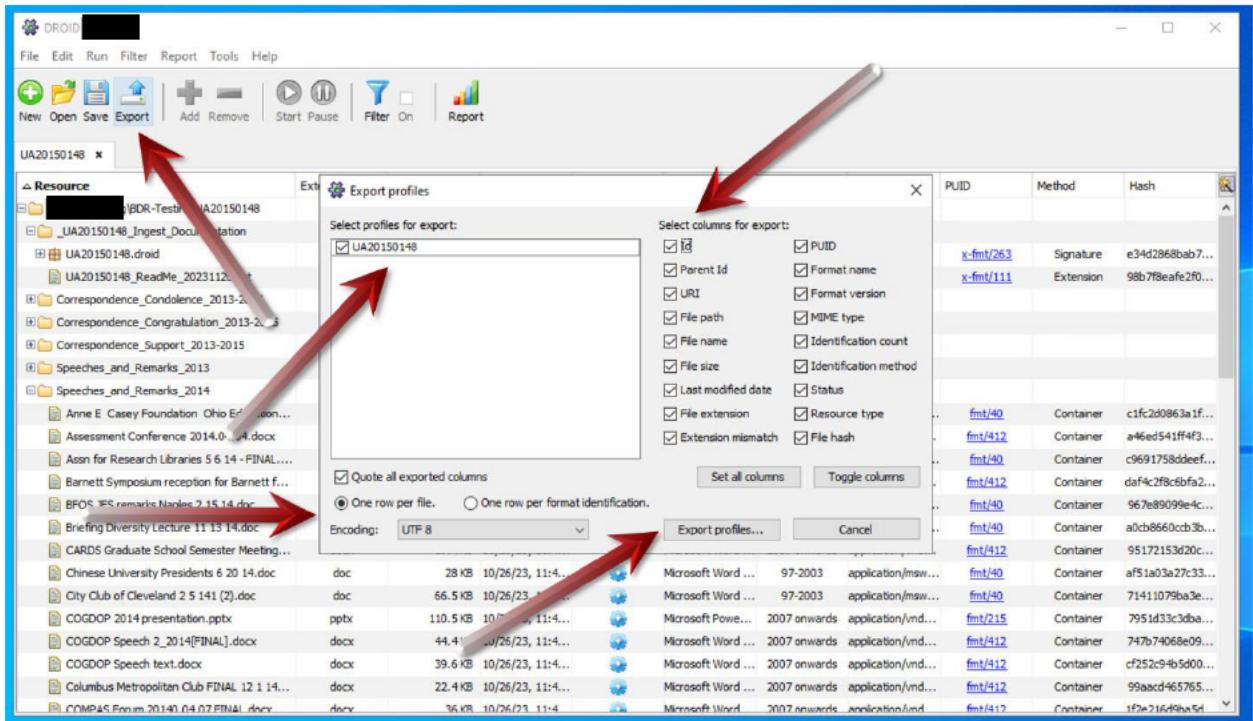


DROID Figure 15c: DROID Report

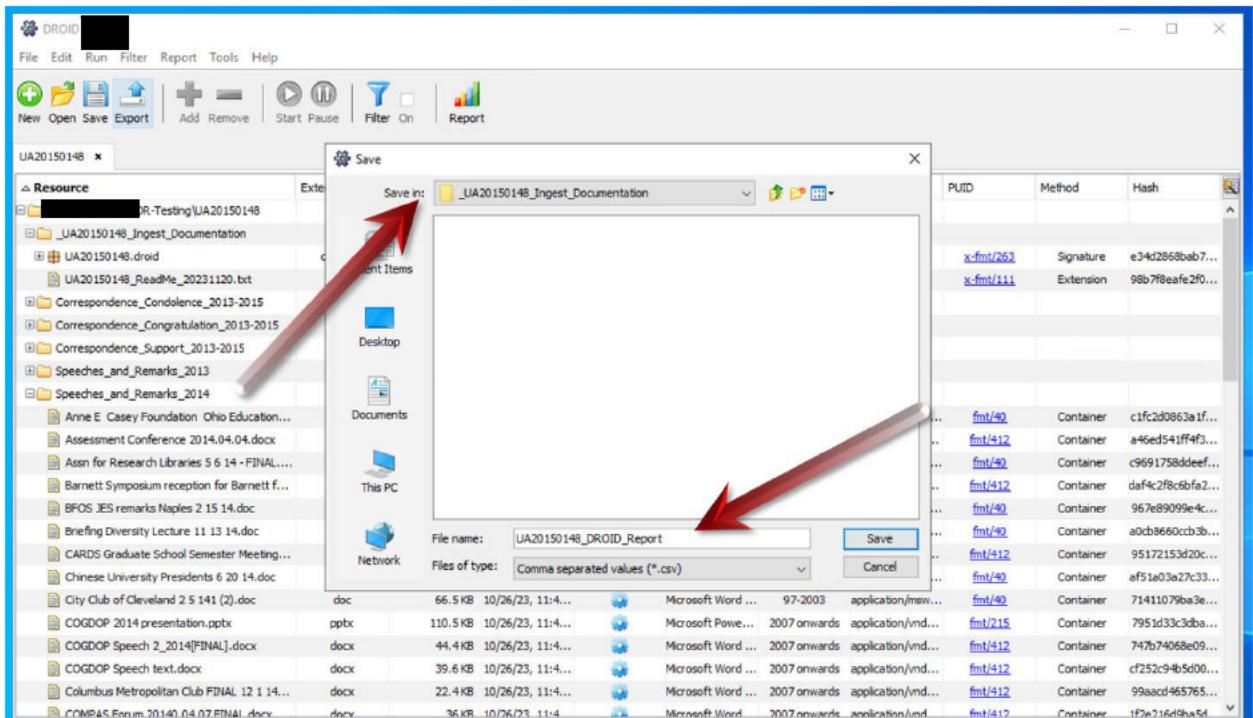


DROID Figure 15d: Saving DROID Report

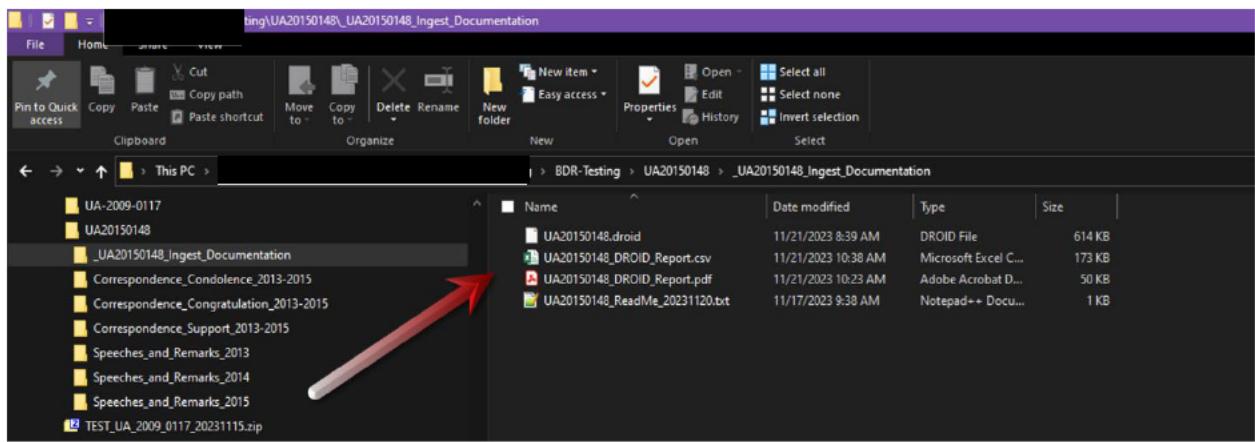
- To export the data, click on the Export icon. This will launch a dialog window.
 - Select your profile
 - Choose which columns to export; we recommend selecting all
 - Make sure the “One row per file” radio button is selected, and Encoding is UTF 8
 - Then click export file
 - Save the file as a .csv, co-locating it in the root directory with the DROID profile and optional DROID report, naming it the same as the DROID profile and accession.



DROID Figure 16a: Manifest export parameters



DROID Figure 16b: Saving manifest export as .csv



DROID Figure 16c: Ingest documentation folder with DROID profile, PDF report, csv manifest and ReadMe file.

PII-Review

This section provides step-by-step instructions to identify potential personally identifiable information (PII) for content that is destined for the Gray Digital Preservation Repository (Gray Repo). The PII that we will be targeting are social security numbers (SSNs) and credit card numbers (CCNs). For University Archives records being accessioned, credit card numbers may include BuckIDs, which are similar to credit cards; however, based upon the University's protocols for handling PII, as well as the age and nature of these records, we are not likely to find BuckIDs among the records. The process described herein provides for the preparation of digitized content (tiff, jpg, image only pdf, etc.) for PII scanning, along with the scanning of text-readable born digital content.

We will utilize a combination of tools to make static image digitized content text-readable via optical character recognition (OCR), and subsequently identify PII. We will use an app, Quick PDF-OCR (aka Fast OCR), developed by Terry Reese, Head of Digital Initiatives that utilizes [Tesseract](#) and [ABBYY FineReader](#) (ABBYY) to [conduct the OCR](#), and Bulk Extractor to [identify PII](#).

While this combination of tools and processes may be run on one's local laptop or workstation, due to current licensing limitations for ABBYY, the University Libraries has setup a dedicated machine for these purposes, which can be accessed via [Remote Desktop](#).

NOTE: If the record set being accessioned and/or transferred to the Gray Repo does not have PDFs or image files, it is unnecessary to conduct the OCR step. Depending upon the size of the accession, one may simply review the file set, or for larger collections review the previously generated DROID report or csv.

Optical Character Recognition (OCR)



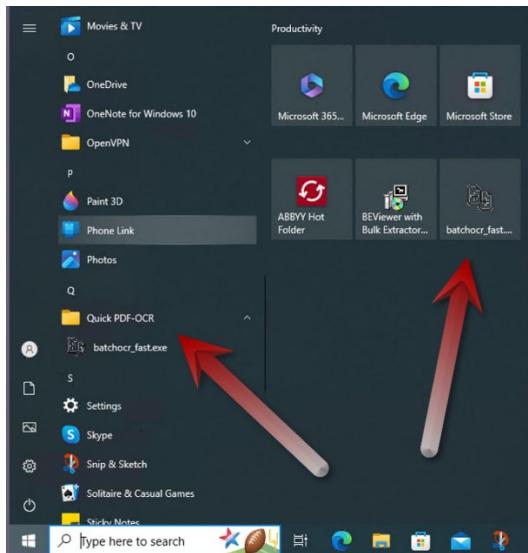
Static image digitized content is not text-readable, therefore the PII identification tool, Bulk Extractor, cannot successfully scan and identify potential PII located within these digital objects.

This may also be true for born digital PDFs of unknown creation processes. We will utilize Quick PDF-OCR to provide a semi-automated approach that allows us to conduct optical character recognition on a set of nested folders, co-locating the resultant output with the original files.

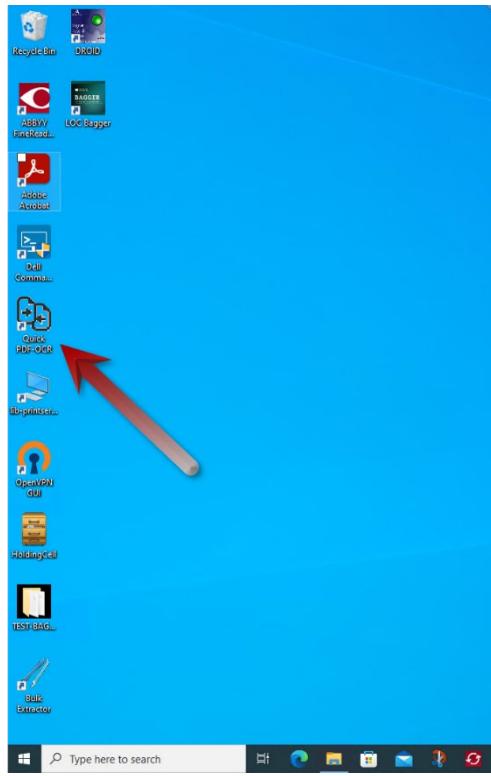
We considered two approaches for OCR output: creation of a searchable text (.txt) file, or a searchable PDF. The searchable text file provides a lighter-weight and smaller footprint within our digital preservation repositories; while the PDF option allows us to produce a redactable ready PDF. We have chosen the former, searchable text option, to be our default; whereas if that initial PII analysis indicates a significant level of potential exposure, we may redo the operation to create the redactable PDFs. Further, if there are compelling reasons to believe the corpus would have the potential for significant exposure/and or a need for searchable PDFs the latter option may be the more prudent choice. Should that be the case please contact the Digital Preservation Department for assistance.

Launch Quick PDF-OCR application

Once you are connected to the [Remote Desktop](#), click on the Windows Menu button, find and open the Quick PDF-OCR application in the Quick-Start menu on the left, or if you have pinned it to the Start-up Tiles (Quick PDF-OCR Figure 1). Alternatively, if you have placed a shortcut on the desktop (renamed from batchocr_fast.exe to Quick PDF-OCR in the below examples), it can be launched from there (Quick PDF-OCR Figure 2).



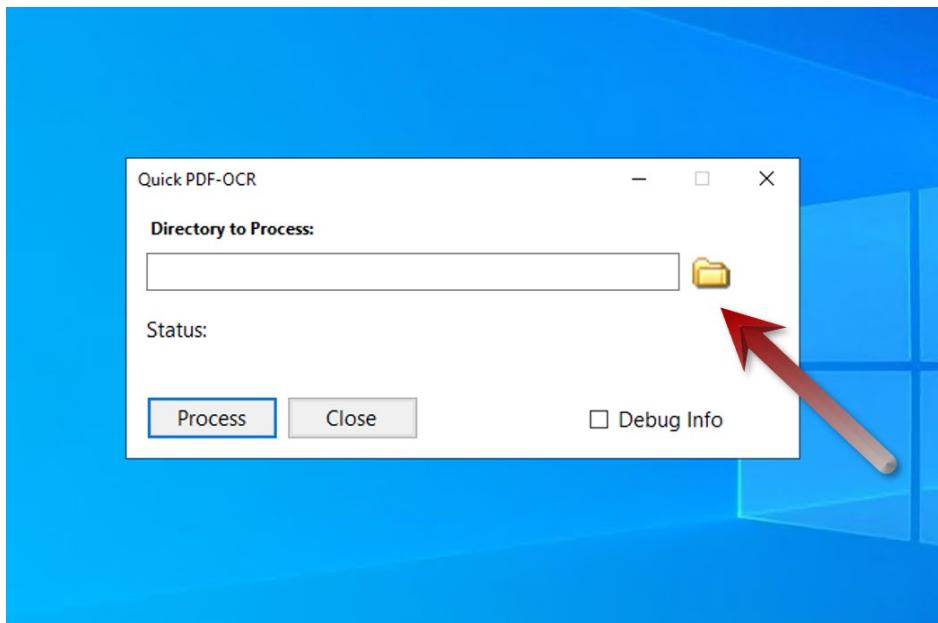
Quick PDF-OCR Figure 1: Desktop with Start launcher and Quick PDF-OCR selected



Quick PDF-OCR Figure 2: Select Quick PDF-OCR (fast_ocr) icon on Desktop

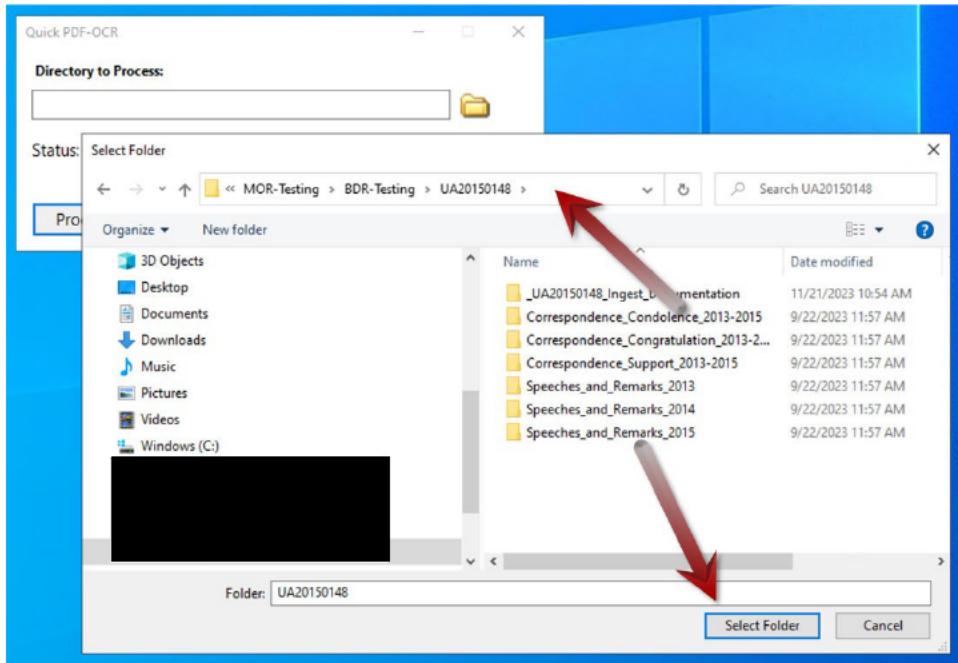
Select content to be OCR'd

- Once Quick PDF-OCR has launched, click on folder icon to locate the content you would like to have processed.



Quick PDF-OCR Figure 3: Folder selection

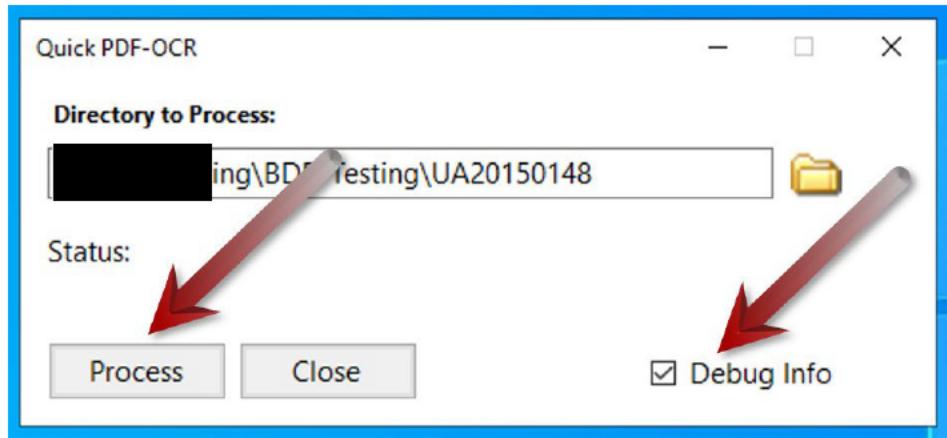
- Navigate through the Windows Explorer interface until you have located the folder with the contents that needs to be OCR'd. Once you have chosen the folder, click the “Select Folder” button, this will bring you back to the Quick PDF-OCR interface.



Quick PDF-OCR Figure 4: Select Folder

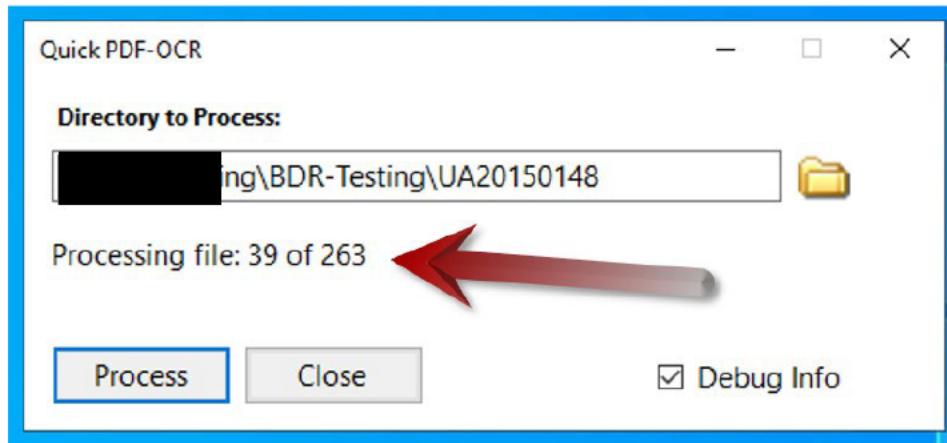
Process Document

- Click on the “Process” button to kick-off the app’s OCR processing.
- Click the Debug Info checkbox. A verbose report will be generated in the dialog box. It is only important if there is an issue. If there is an issue, an actual debug_file.txt will be created in the root folder, documenting what could not be OCR'd.



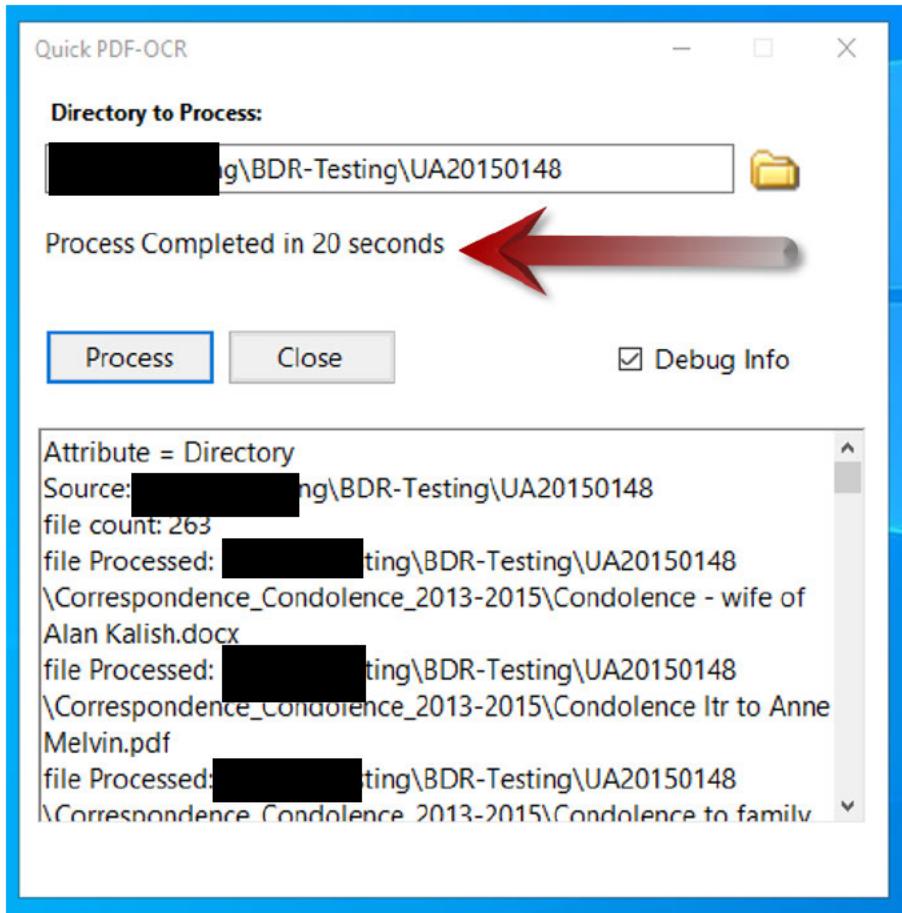
Quick PDF-OCR Figure 5: Start Process

- The app will now show progress of “Processing file # of #”



Quick PDF-OCR Figure 6: Processing progress indicator

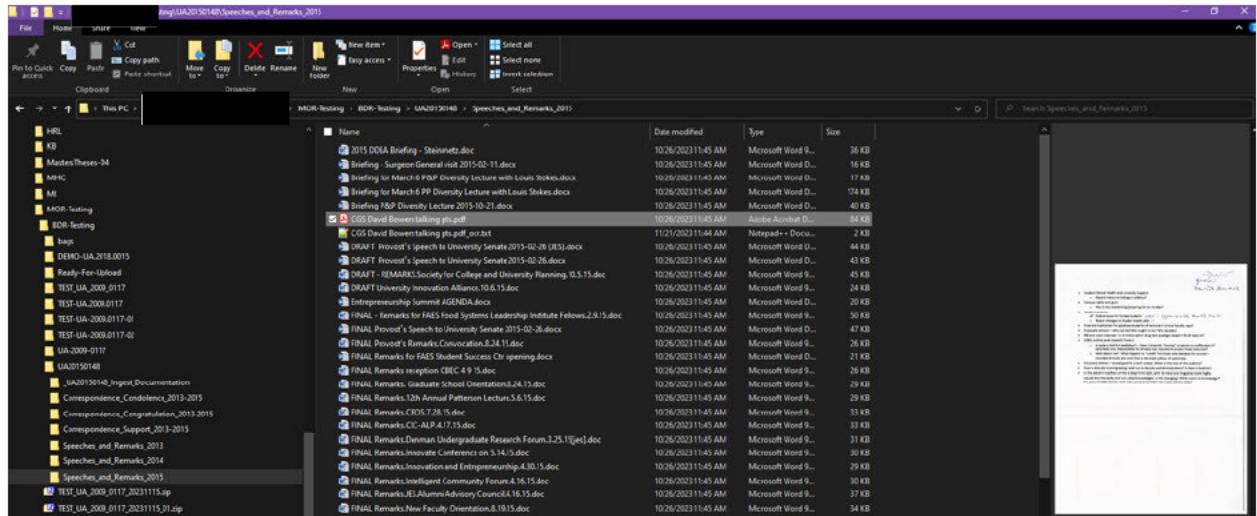
- When the process is done the app will indicate “Process Completed”. Depending upon the number of files and file sizes, the time to complete will vary. Should an error occur prior to the process completing, please contact the Digital Preservation Department as soon as possible.



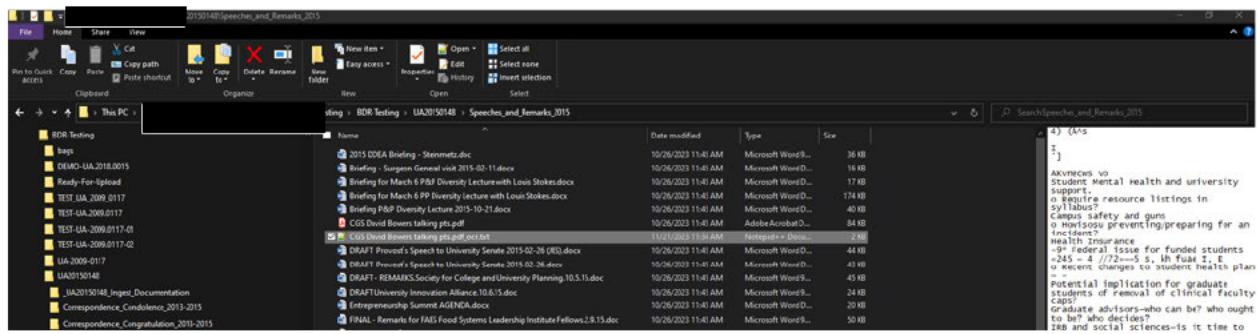
Quick PDF-OCR Figure 7: Process complete

Results

- The text files created for the OCR'd PDF and image files, will be co-located with your original files. Only PDFs and images where text was recognized will have the new companion text file. The new .txt files will consist of the original file name with an “_ocr” suffix (e.g. “CGS David Bowers talking pts.pdf” and “CGS David Bowers talking pts.pdf_ocr.txt”).

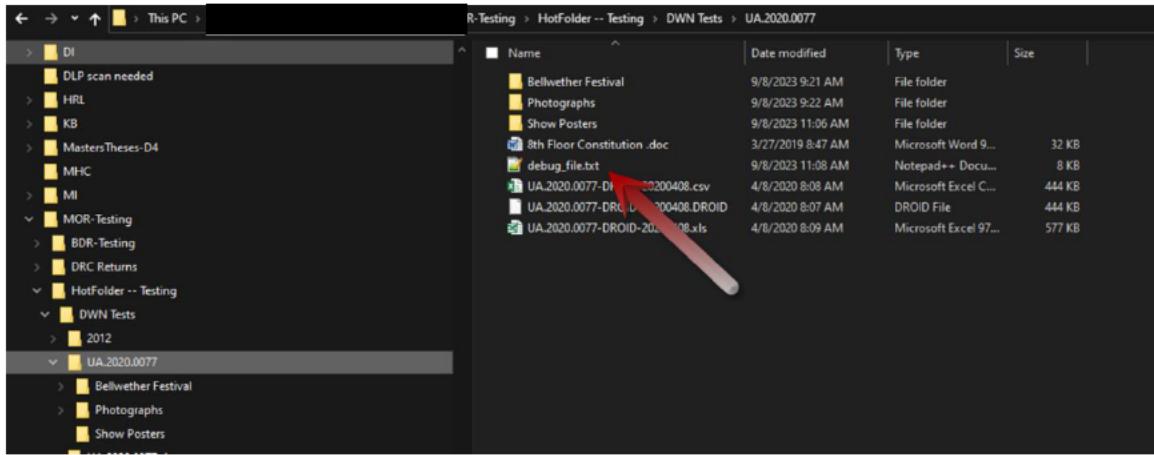


Quick PDF-OCR Figure 8: CGS David Bowers talking pts.pdf in Windows File Explorer viewer



Quick PDF-OCR Figure 9: CGS David Bowers talking pts.pdf_ocr.txt

- As previously noted, in addition to the companion OCR'd text files, Quick PDF-OCR may also create a log file called “debug_file.txt” that will be located in the root folder. This log indicates files that could not be examined. These files should be investigated further by the curatorial staff, potentially in consultation with the Digital Preservation Department. If the debug_file.txt is generated, it should be moved to the _AccessionID_Ingest_Documentation folder and renamed AccessionID_debug_file.txt. If no debug_file.txt is generated, it is best to note it in the ReadMe file.



Quick PDF-OCR Figure 10: "debug_file.txt" located in the root folder analyzed

```

1 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49231465_10156364360987879_4767736419899893760_o_101563643609462875.jpg
2 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49421969_101563643452875_665248471019311616_o_10156364361442875.jpg
3 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49564156_1015636434362057875_82285261015636434362047875_.jpg
4 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49646446_1015636434362057875_82285261015636434362047875_.jpg
5 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49651129_1015636434362057875_82285261015636434362047875_.jpg
6 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49621921_10156364355512075_572515019229346204_o_10156364355412075.jpg
7 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49631316_10156364355512075_572515019229346204_o_10156364355412075.jpg
8 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49671103_10156364360232075_3650863519742034680_o_10156364360222875.jpg
9 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49671743_10156364360232075_3650863519742034680_o_10156364360222875.jpg
10 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49602074_10156364361017875_6824366157584138240_o_101563643609597875.jpg
11 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49600559_10156364361007075_0779421913960231264_o_101563643609592075.jpg
12 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49605604_10156364359767875_8649795849011855360_o_10156364359727875.jpg
13 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49834039_10156364361697875_6381084711857691489_o_10156364361692875.jpg
14 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49831957_10156364359412875_2990042315416666112_o_10156364359312875.jpg
15 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\49840161_1015636435912875_941427694225063939_o_10156364359157875.jpg
16 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\50018094_1015636435912875_4591446539095769089_o_10156364359147875.jpg
17 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\50018094_1015636435912875_4591446539095769089_o_10156364359147875.jpg
18 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\501330176462_10156364362552875_89198846562123520_o_10156364362532875.jpg
19 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\50245895_1015636436042875_892437815229298689_o_1015636436002875.jpg
20 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Emotional Breakdown Show - Nov 2018\50272105_10156364360552975_58737261054721024_o_1015636436052875.jpg
21 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7445.JPG
22 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7446.JPG
23 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7451.JPG
24 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7471.JPG
25 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7480.JPG
26 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7487.JPG
27 Testing\DWNN Tests\UA.2020.0077\Photographs\Shows\Murder Mystery Show - March 2019\IM6_7511.JPG
28

```

Quick PDF-OCR Figure 11: "debug_file.txt" log details

- At this point, barring any other investigation of files based upon the "debug_file.txt" you should be ready to begin the PII analysis.

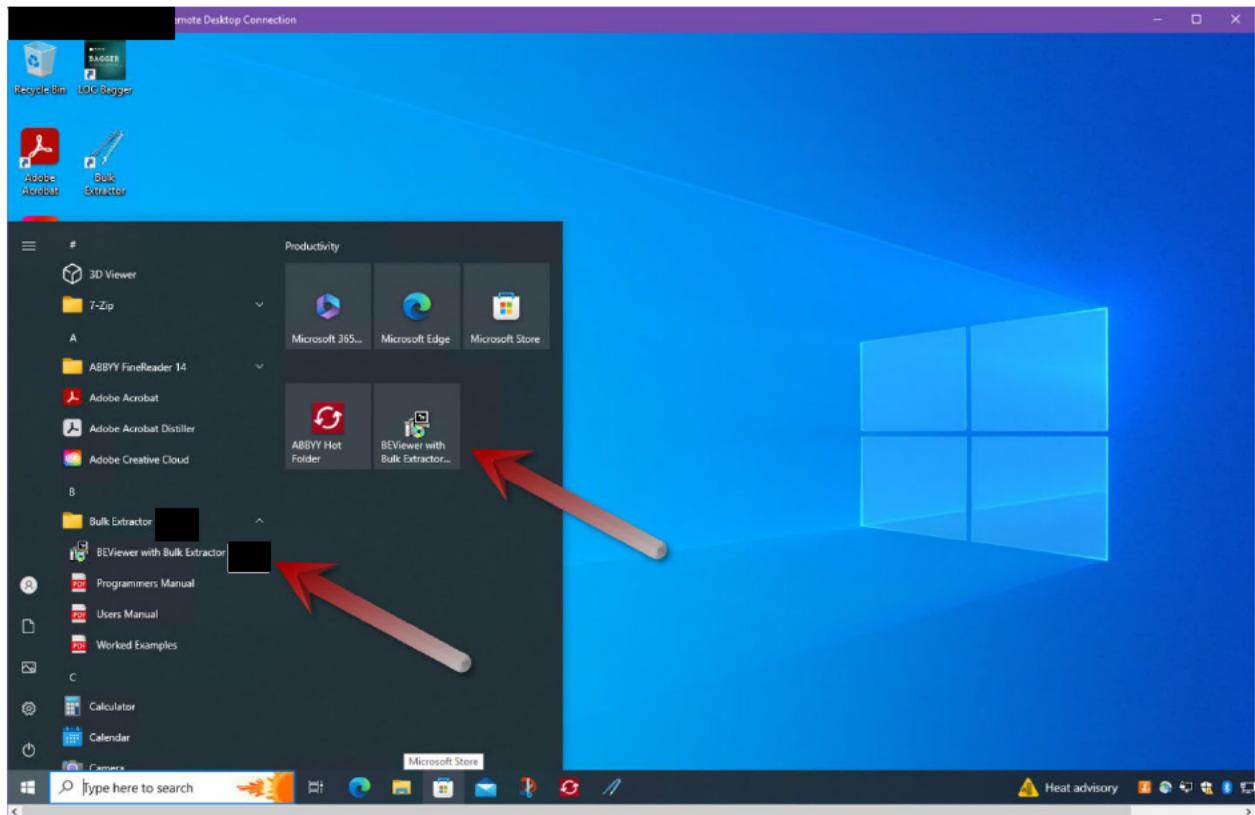
Identifying PII Data with Bulk Extractor



Bulk Extractor, like Quick PDF OCR and ABBYY Fine Reader has been installed on the [Remote Desktop](#) workstation.

Launch Bulk Extractor

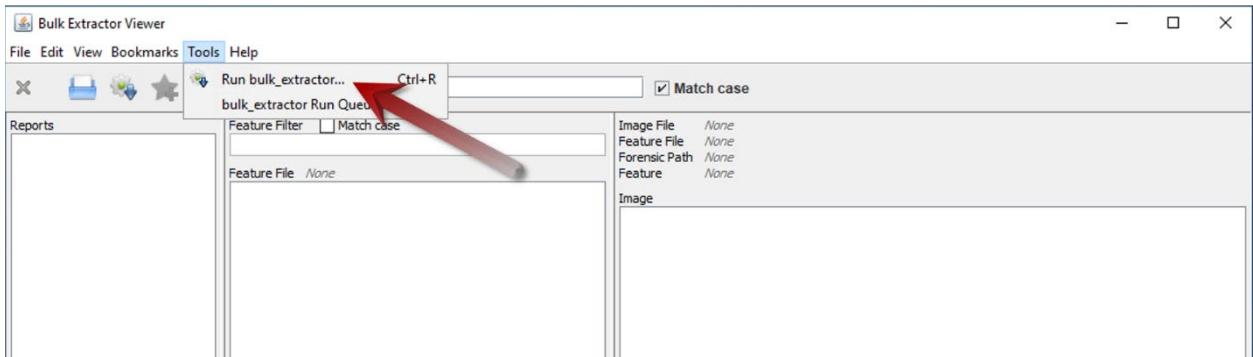
Once you are connected to the Remote Desktop, click the Windows Start Menu button, find and open the Bulk Extractor application in the Quick-Start menu on the left, or if you have pinned it to the Start-up Tiles.



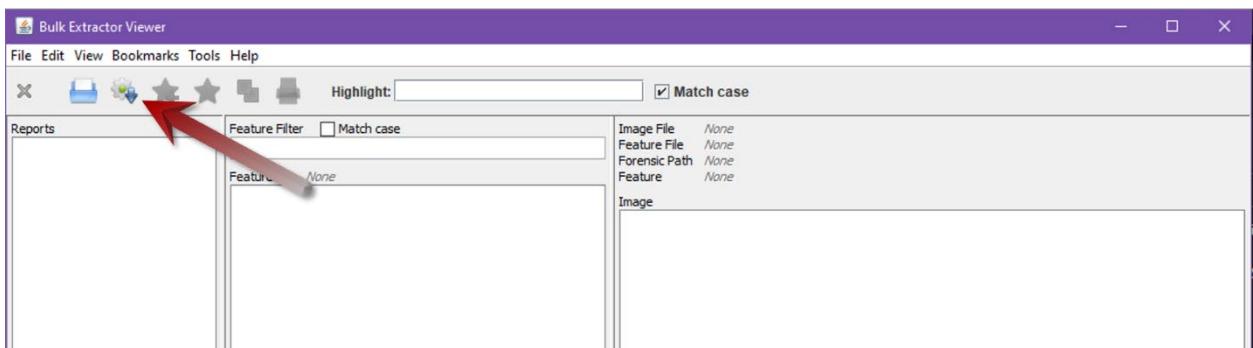
Bulk Extractor Figure 1: Desktop with Start launcher and Bulk Extractor selected

Run bulk_extractor

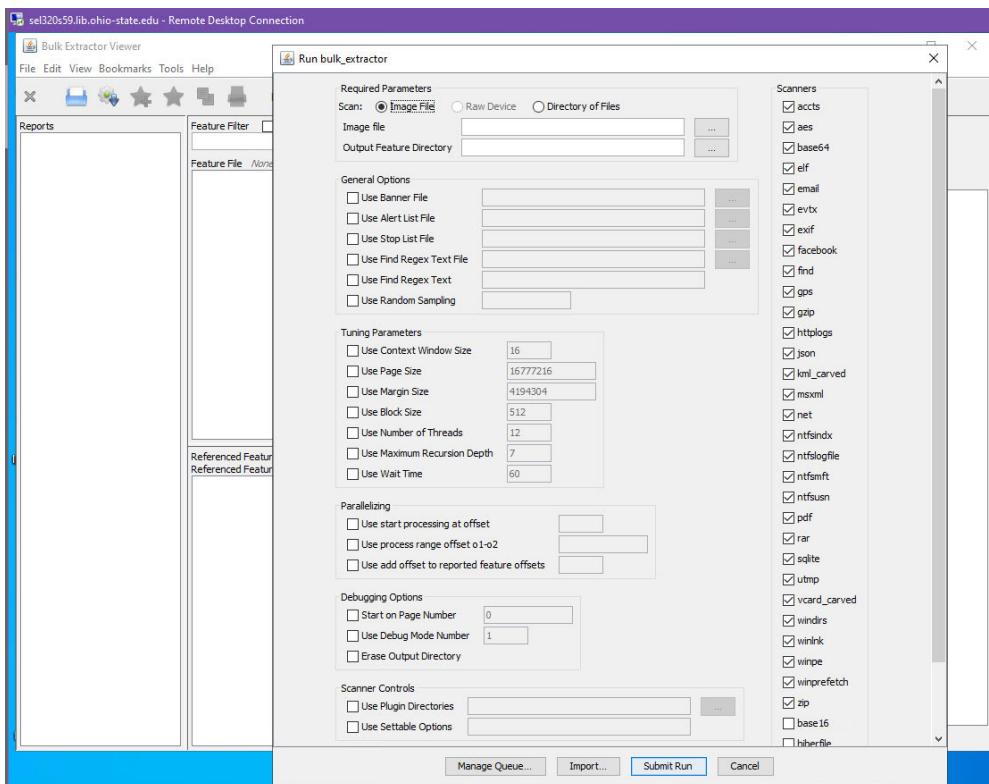
- Select “Tools” and then “Run bulk_extractor” (Bulk Extractor Figure 2a); alternatively, you can click on the gear with down arrow next to the printer icon in tool bar (Bulk Extractor Figure 2b). This will bring up the Bulk Extractor Parameters window, which will likely need to be resized (Bulk Extractor Figure 3).



Bulk Extractor Figure 2a: Run bulk_extractor via Tool pulldown menu



Bulk Extractor Figure 2b: Run Bulk Extractor via gear/arrow icon

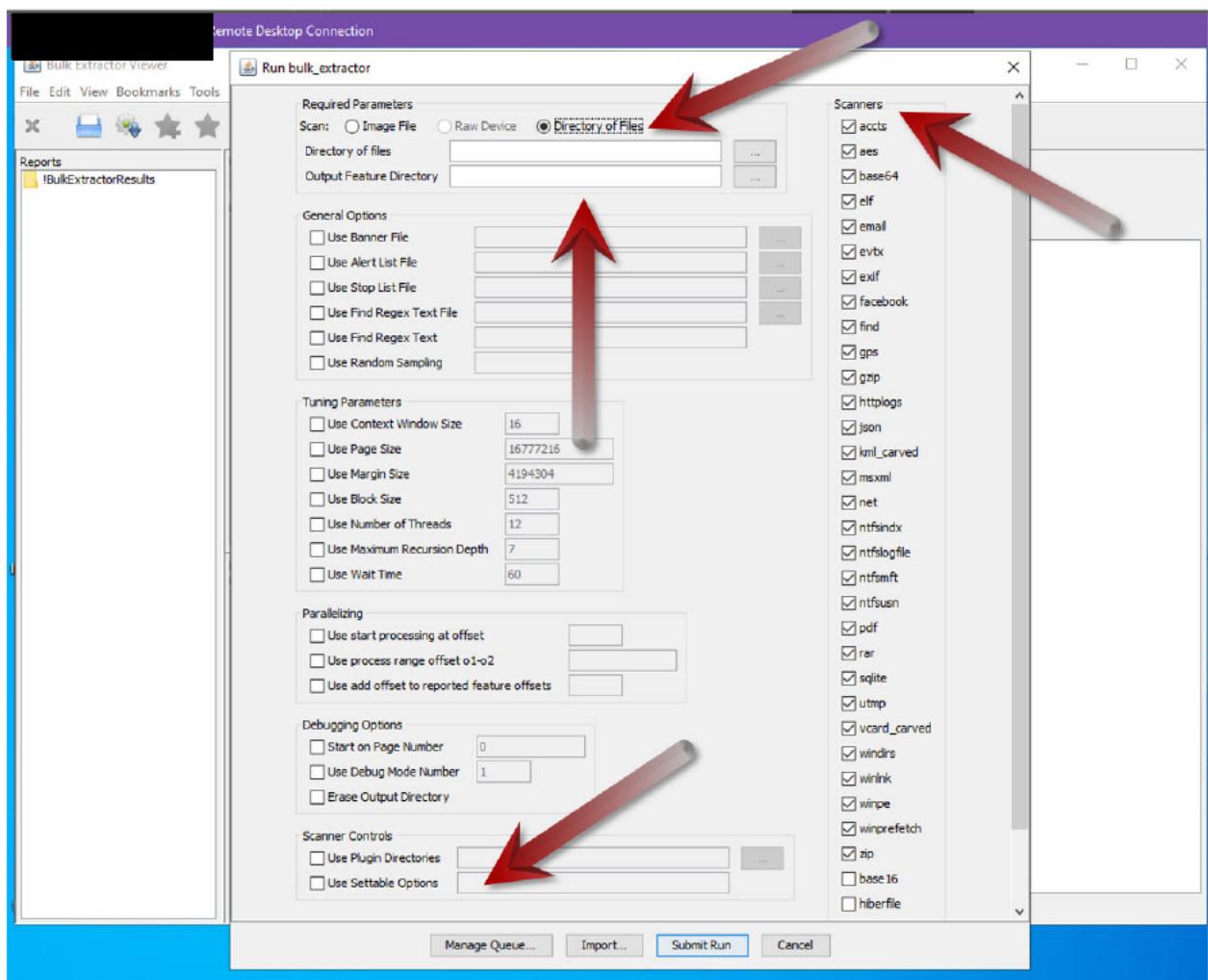


Bulk Extractor Figure 3: Bulk Extractor Parameters

Required Parameters

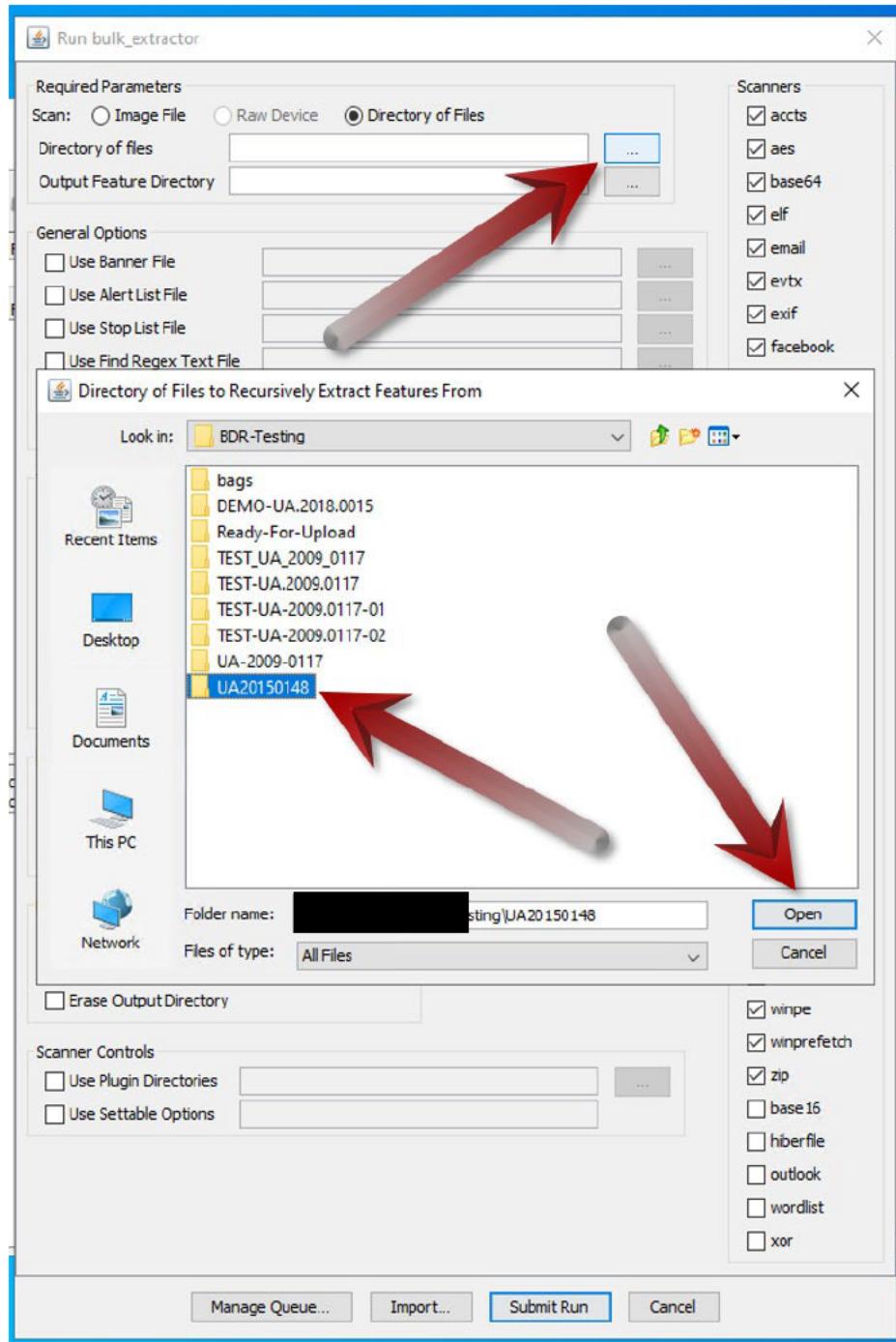
You will need to set the following parameters (Bulk Extractor Figure 4):

- Scan: Directory of Files (Click the “Directory of Files”)
 - Directory of Files
 - Output Feature Directory
- Scanners
- Scanner Controls: User Settable Options



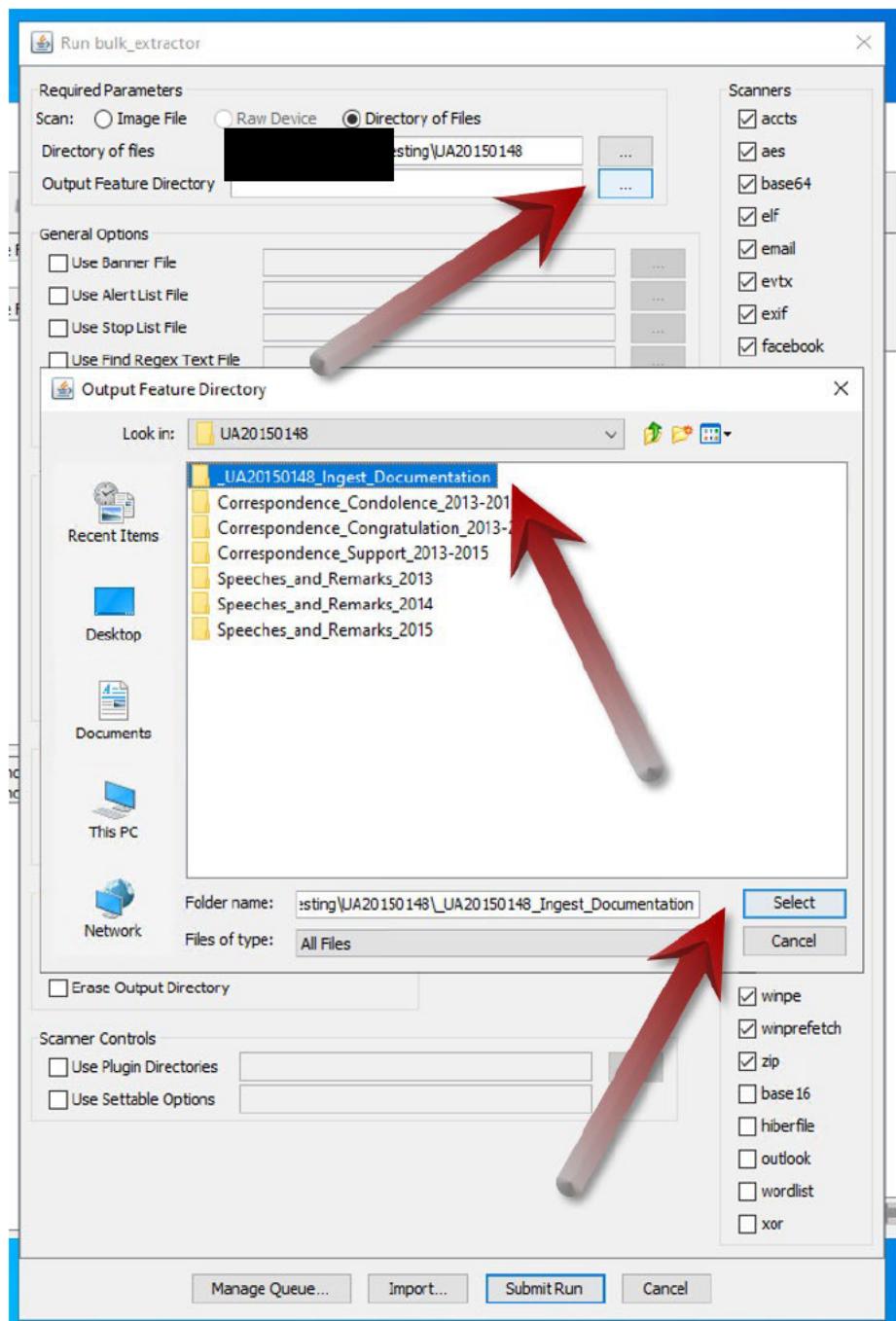
Bulk Extractor Figure 4: Setting Bulk Extractor Parameters

- Directory of Files (Bulk Extractor Figure 5):
 - Click on the ellipsis “...”
 - Navigate down to the source directory that needs to be scanned
 - Click “Open”



Bulk Extractor Figure 5: Selection of Directory of Files

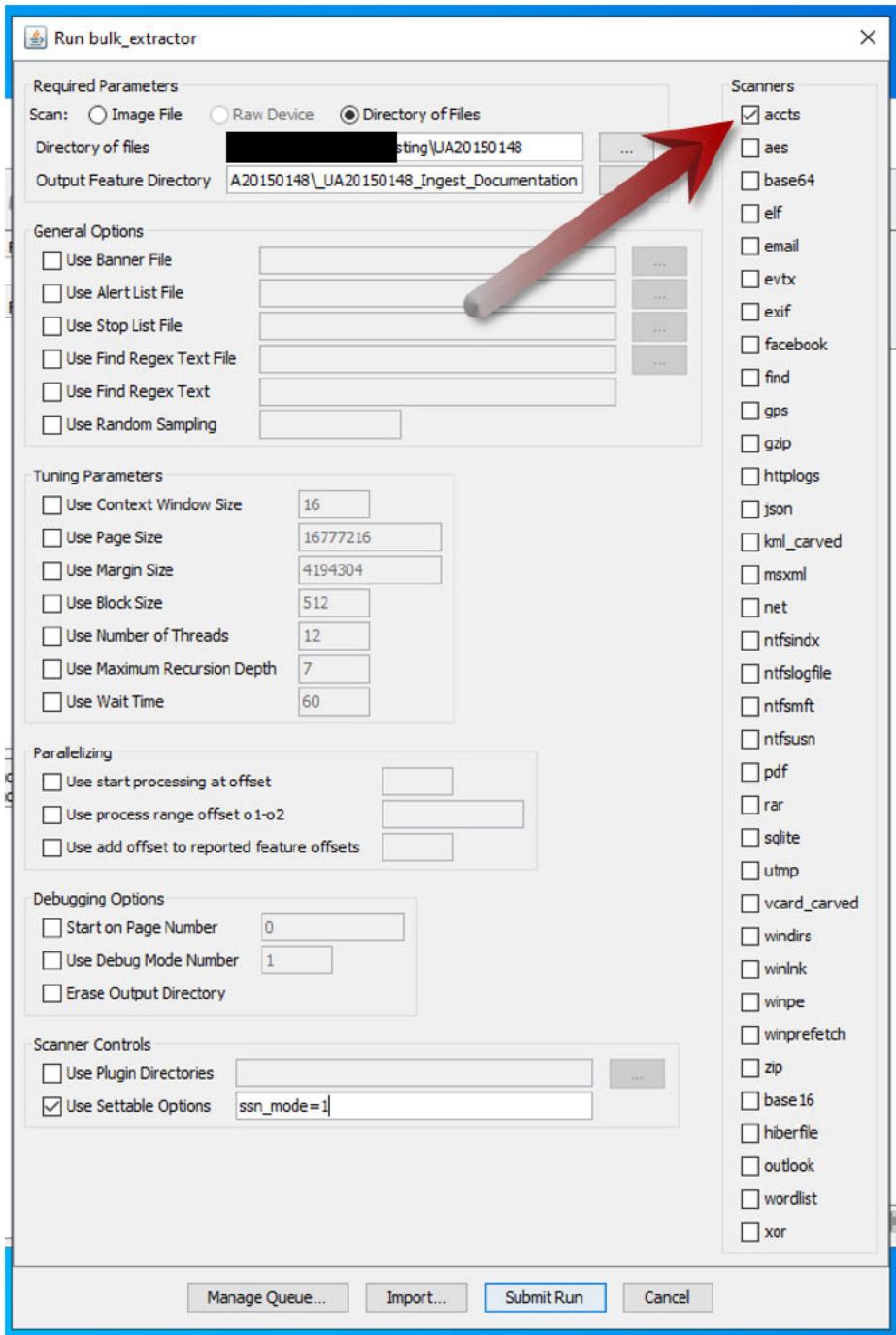
- Output Feature Directory (Bulk Extractor Figure 6):
 - Click on the ellipsis “...”
 - Navigate down until the source directory that needs to be scanned is opened
 - Select the _AccessionID_Ingest_Documentation folder
 - Click “Select”



Bulk Extractor Figure 6: Output Feature Directory

Scanners

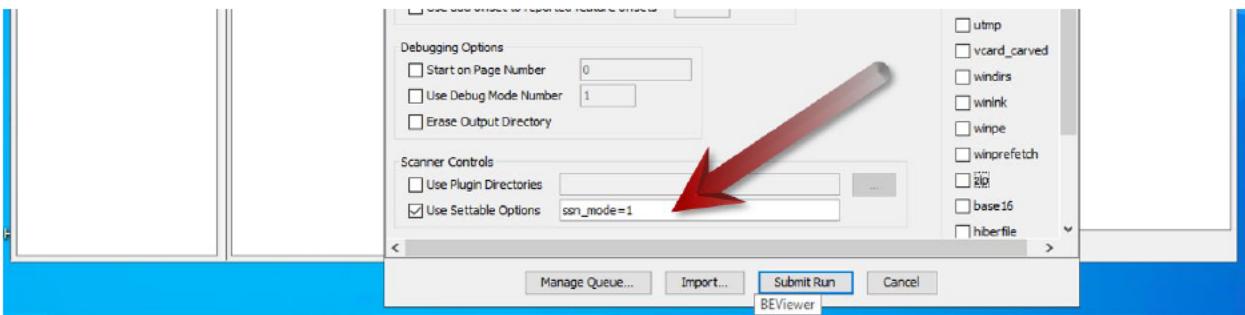
This process will be a little tedious, but you need to unclick all scanners except for "accts"



Bulk Extractor Figure 7: Scanner Selection

Scanner Controls

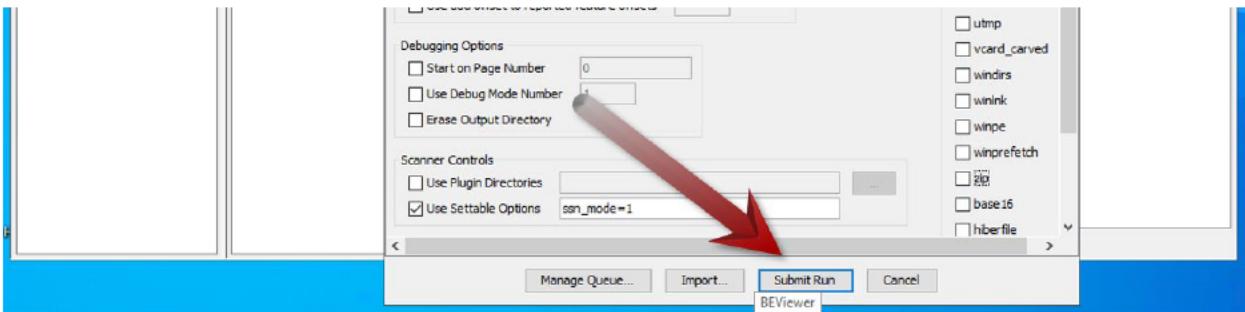
Located near the bottom of the window, in "Scanner Controls" check "Use Settable Options" and enter "ssn_mode=1" in the text box. This allows for more nuanced results when searching for potential social security numbers (SSNs) by identifying and bypassing false-positive results.



Bulk Extractor Figure 8: Scanner Controls – Use Settable Options

Submit Run

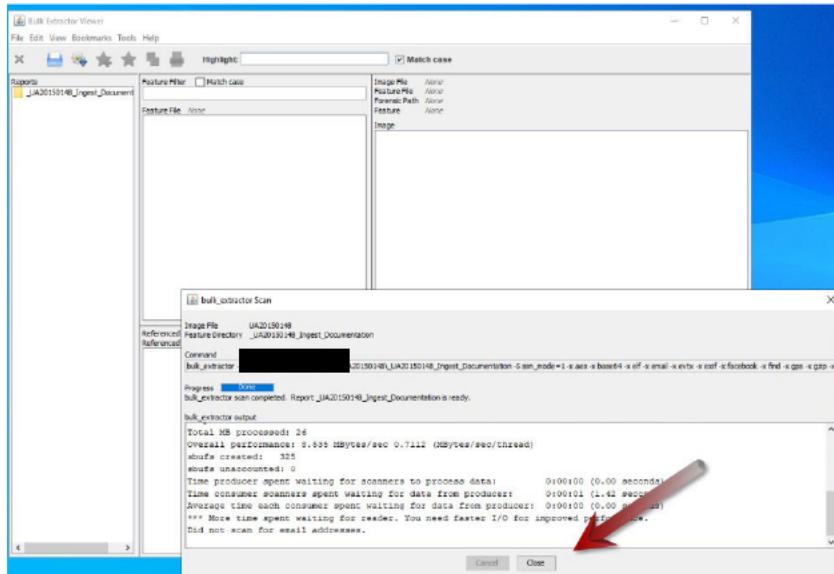
Finally, you are ready to have Bulk Extractor search for PII. Click “Submit Run”



Bulk Extractor Figure 9: Submit Run

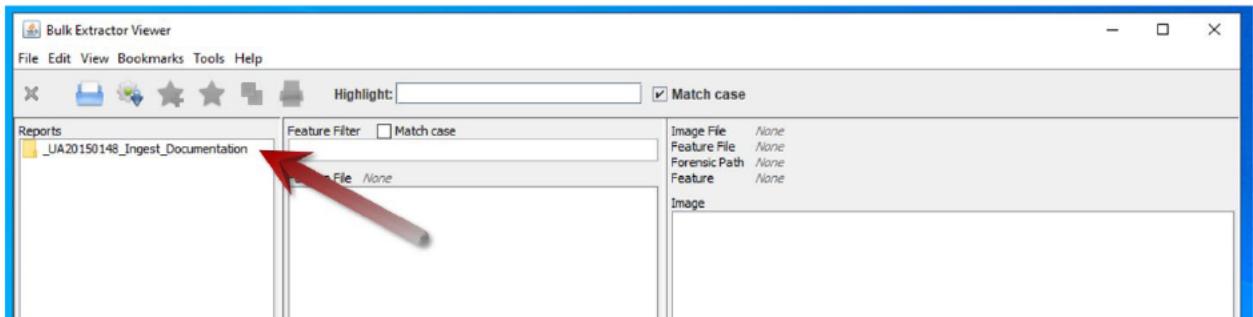
Results

When the process has completed its run, click the Close button.

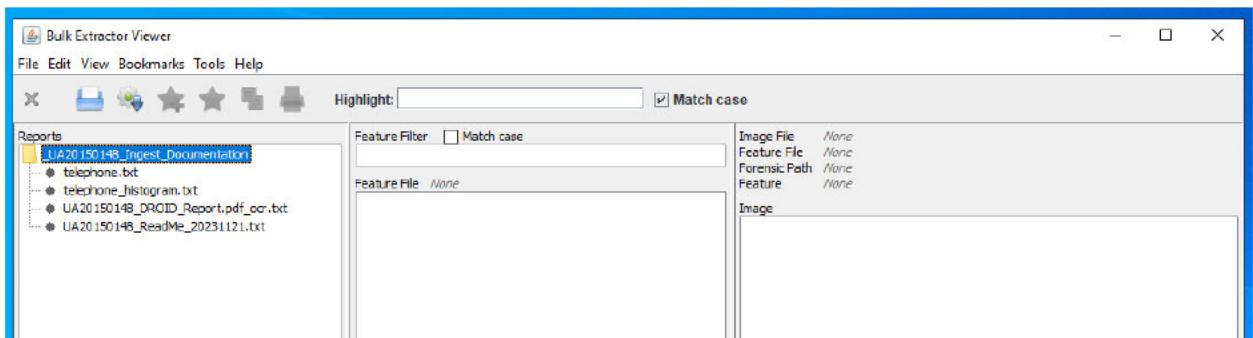


Bulk Extractor Figure 10: Bulk Extractor Completed Run

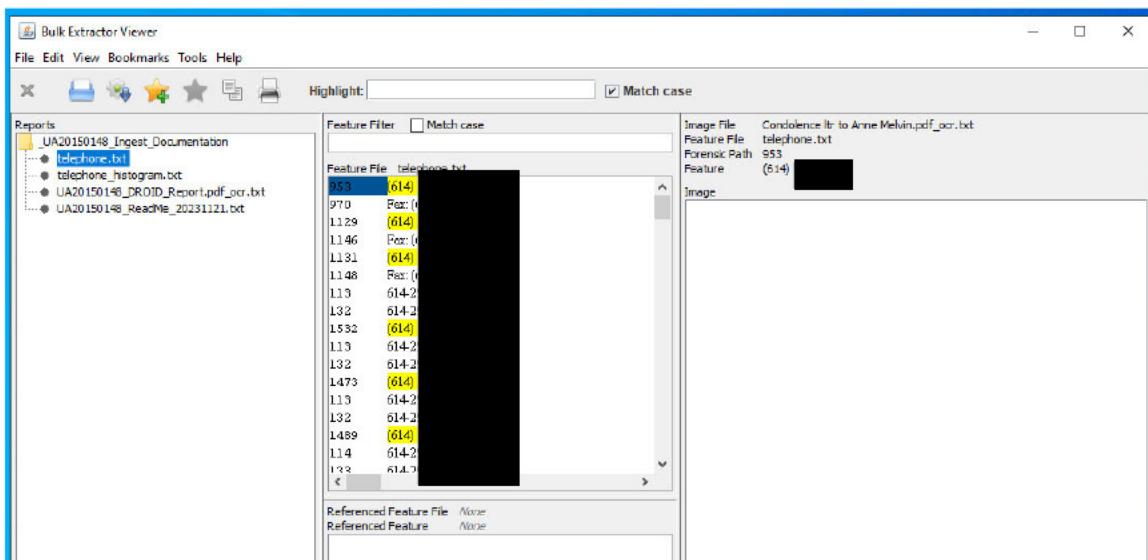
Reports for the data that are discovered are loaded in the left pane; click on the folder (Bulk Extractor Figure10) to see which reports have been generated that include discovered data (Bulk Extractor Figures 11 and 12). If any PII data is found, it will show up in a pii.txt report. In this particular example no PII was found, therefore no PII report is shown. However, Bulk Extractor did identify telephone numbers.



Bulk Extractor Figure 11: Results Folder

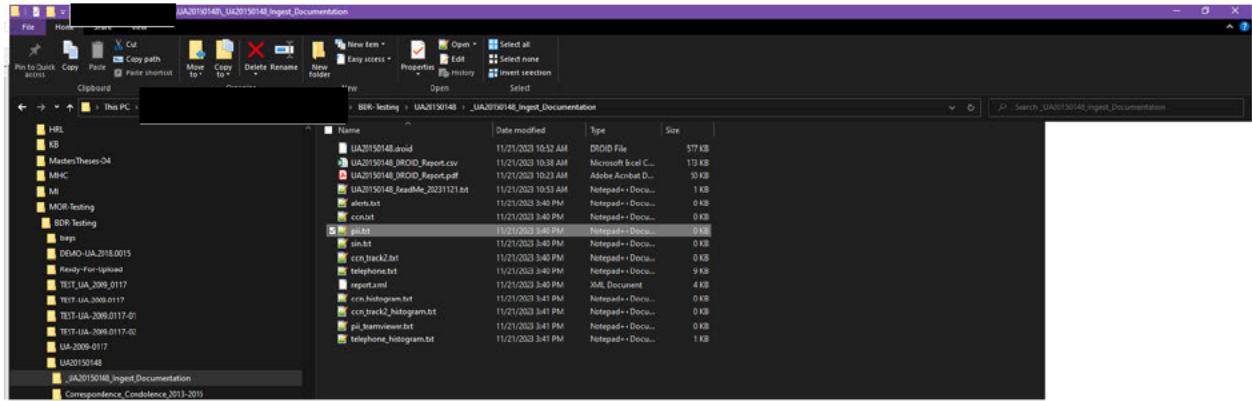


Bulk Extractor Figure 12: Reports along with existing .txt files that have data in _AccessionID_Ingest_Documentation folder



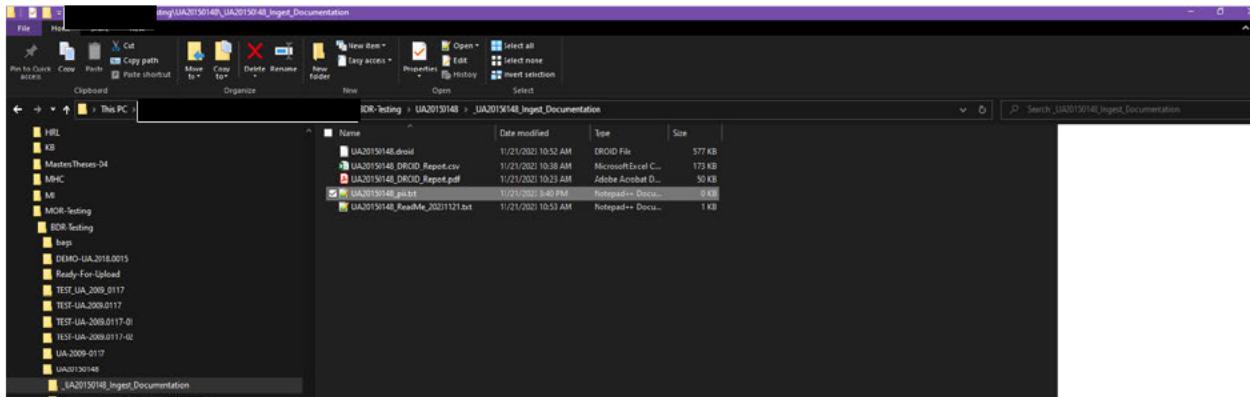
Bulk Extractor Figure 13: telephone.txt scanner report detail

In all, Bulk Extractor created ten (10) reports, albeit they have no data except for the telephone and telephone histogram reports. An additional file, report.xml is created in order to render data in the Bulk Extractor interface.



Bulk Extractor Figure 13a: List of Bulk Extractor generated reports with pii.txt highlighted

The only report we need to retain is the pii.txt, regardless of whether PII was identified. It will need to be bagged up with the archival content, DROID manifest files and ReadMe file, as well as retained in the Gray Repo's Administrative Team. We strongly encourage adding a note to the ReadMe file confirming no PII was found. Should there be a curatorial reason for retaining other reports that generated data, it should be documented in their processing notes and ReadMe file. If retained, those reports should also be bagged for ingest, and maintained in Gray Repo's Administrative Team.



Bulk Extractor Figure 13b: File list after renaming and deletions

Bagging

This section provides step-by-step instructions to package digital content that is destined to be uploaded into the Gray Digital Preservation Repository (Gray Repo or Gray Repo) using the [Library of Congress' Bagger](#) software. Bagger uses the BagIt specification with a graphical user interface (GUI) rather than command line input. This version will also allow for individualized profiles to be created, applying specific and consistent metadata about the bagged content, based on the uploading needs of the content stewards.

Bagger is a Java based software, and therefore Java needs to be installed on one's local laptop or workstation to run it. Due to the time and effort required to install this on each computer, the University Libraries has set up a dedicated machine for this purpose, which can be accessed via [Remote Desktop](#).

Bagger/BagIt

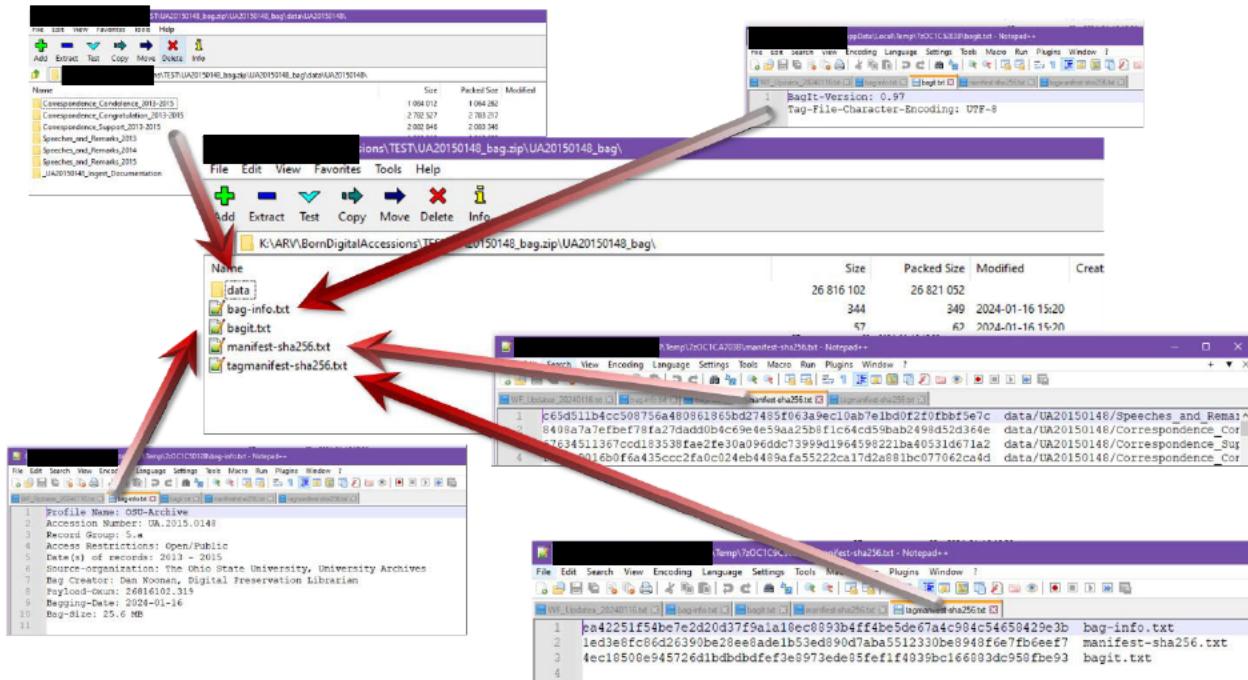


The BagIt specification and Bagger software are designed to support the packaging, storage, and transfer of digital content using a “bag” that is a single top-level folder that includes the folders/files to be transferred and stored in the Gray Repo, along with metadata created during the bagging process. It can be created as a standard folder or in a compressed format such as ZIP or TAR. At this point in time, we will not be compressing the files.

The main components of a bag are the (see Bagger Figure 1):

- top level folder
- payload “data” folder which contains all the files and folders being bagged
- bag metadata files:
 - bag-info.txt, which includes the metadata created within the profile along with the payload oxum (a combination of the overall file size and number of files) and the bag size
 - bagit.txt, which indicates what version of BagIt was used along with the character encoding
 - manifest-sha256.txt, which included a file-by-file manifest of all files within the data folder along with a checksum;
 - tagmanifest-sha256.txt, which includes checksums for the bag-info.txt, bagit.txt and manifest-sha256.txt files

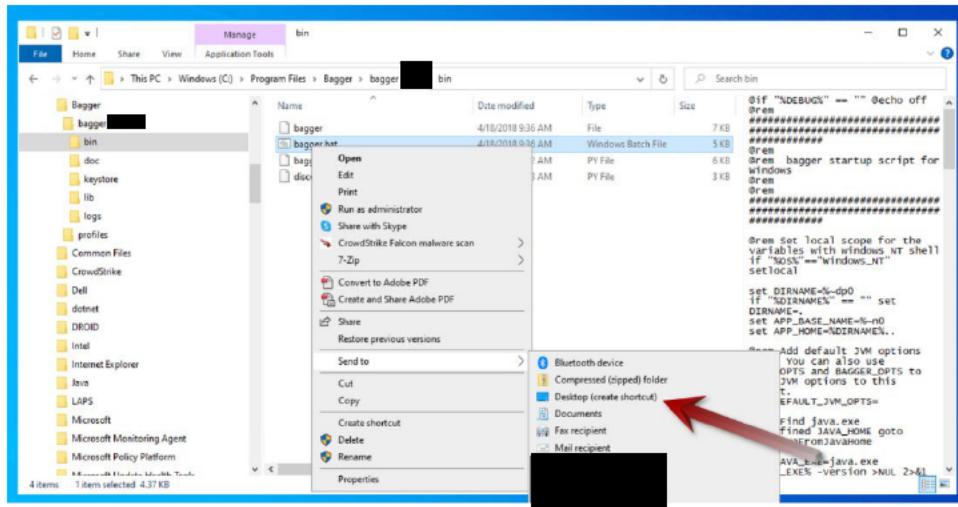
If you compare this process to physical material, think of this bag as the cubic foot box, the metadata as the label that you place on the outside the box, as well as a file/folder index placed within the box, which all together allows for quick and easy determination of what is inside that box.



Bagger Figure 1: Bag Components

Shortcut

- Navigate to C:\Program Files\Bagger\bagger-2.8.1\bin
- While you can just “Double Click” on bagger.bat file to launch the program, it will be easier in the long run to have a desktop shortcut
- “Right Click” on the bagger.bat file
 - Click on “Send to”
 - Click on “Desktop (create shortcut)”



Bagger Figure 2: Create desktop shortcut

- a desktop shortcut “bagger.bat – Shortcut” will be created
- You can rename the shortcut to just “Bagger” or “LOC Bagger” as we did, and change its icon; contact the Digital Preservation Department for assistance.



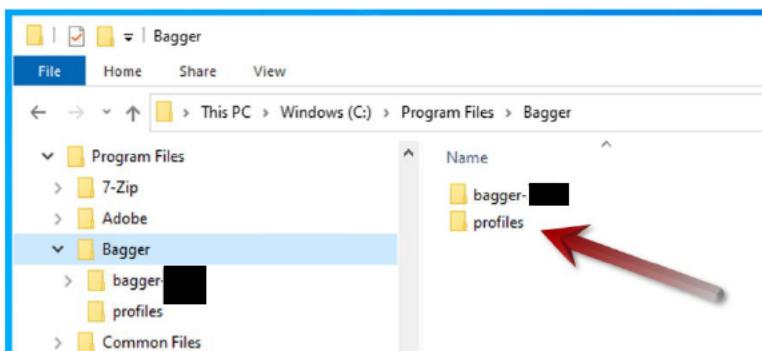
Bagger Figure 3: Bagger desktop shortcut example

Setup Bagger Profile

Before you start using Bagger for the first time, there is a little bit of setup that needs to be done on the part of the user. Bagger allows the use of metadata profiles to be created that are ultimately rendered in the bag-info.txt as discussed above. The profile(s) can be individually tailored to the needs of the individual collecting unit(s). By utilizing a metadata profile that is tailored to your unit, entering the appropriate amount of metadata will be streamlined. Your unit needs to collaborate with the Digital Preservation Department to define your particular profile.

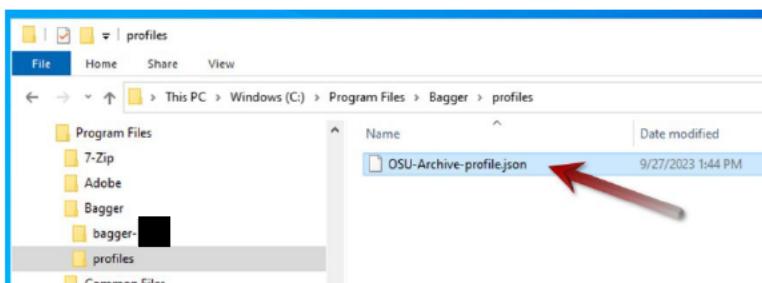
The following steps will guide you through the process of making sure that your correct profile loads when you start up the Bagger application:

- Navigate to C:\Program Files\Bagger
- Open the Profiles folder



Bagger Figure 4: Bagger profiles folder

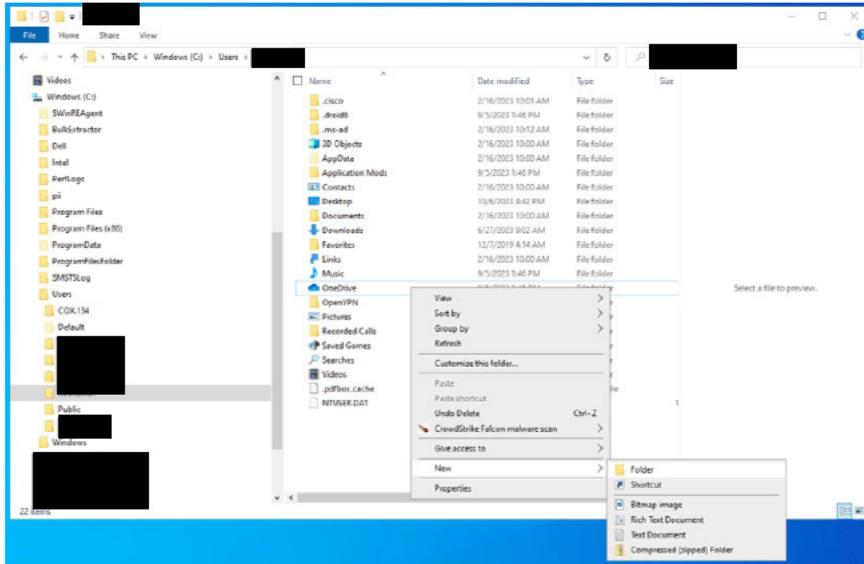
- Select the unit appropriate file (eg. OSU-Archives-profile.json) and Copy it.



Bagger Figure 5: Example of Bagger profile .json file

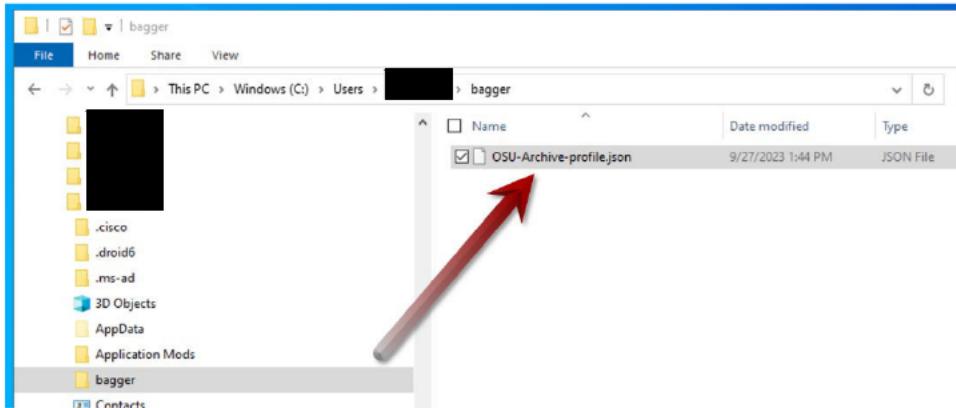
- Next, navigate back to your user folder (e.g. C:\Users\[REDACTED]).

- In your user folder, create a new folder (right click the mouse button, highlight New, and select Folder) within it called “bagger”; this folder name is case sensitive, and must be lowercase.



Bagger Figure 6: Creating "bagger" folder in Windows User profile

- Go into the new “bagger” folder and Paste the previously copied profile into it.



Bagger Figure 7: Pasting unit appropriate json profile in "bagger" folder

NOTE: You only need to do this step once prior to using Bagger for the first time. You will never need to do this again on subsequent uses unless you need to change or add profiles in the future.

Run Bagger

Bagger should now be ready to use. Your accession should be ready to be bagged. This bag will contain the:

- folder(s) and file(s) that are being accessioned
- [AccessionID_Ingest_Documentation](#) folder and files

NOTE: Do not continue through these steps if one of those elements is missing.

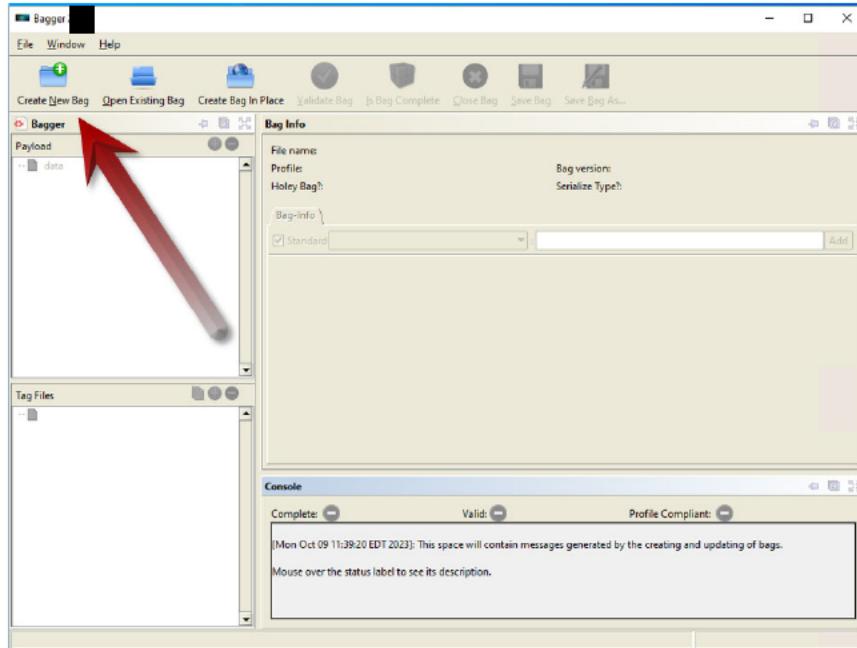
- Launch Bagger by clicking on the desktop shortcut.



Bagger Figure 8: Launch Bagger

Create a Bag

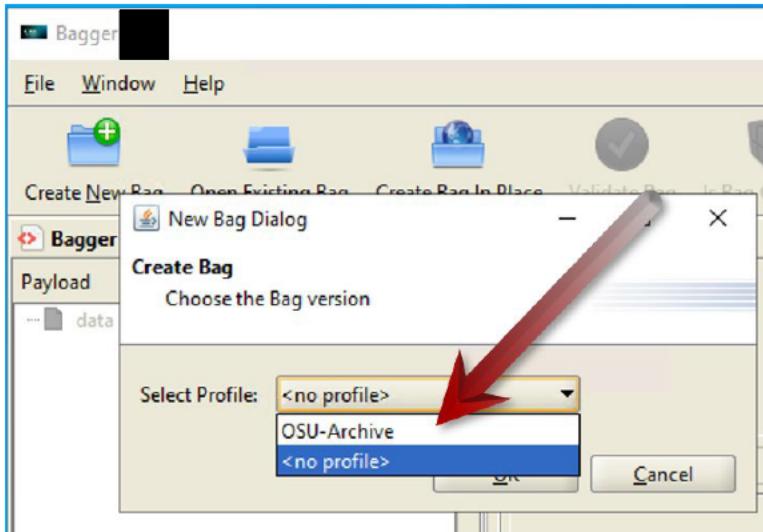
- Select “Create New Bag” in the upper left corner of the application.



Bagger Figure 9: Create New Bag

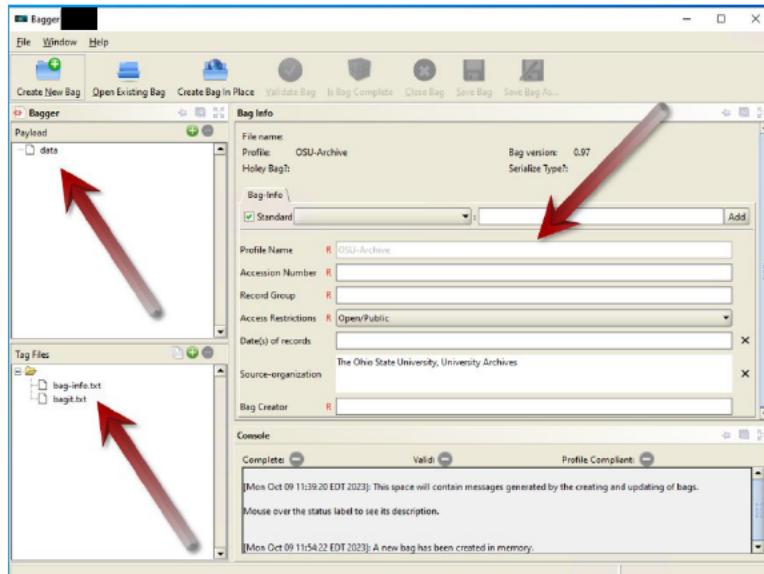
- This opens a dialog window asking you to select a profile. Click the down arrow to show the options. Presuming the prior step of including the appropriate up the profile was successful, you should only see one option, which is the profile

tailored to your specific use (e.g. OSU-Archive). If not, please contact the Digital Preservation Department for assistance. Select the profile and click OK.



Bagger Figure 10: Selecting appropriate profile for new Bag

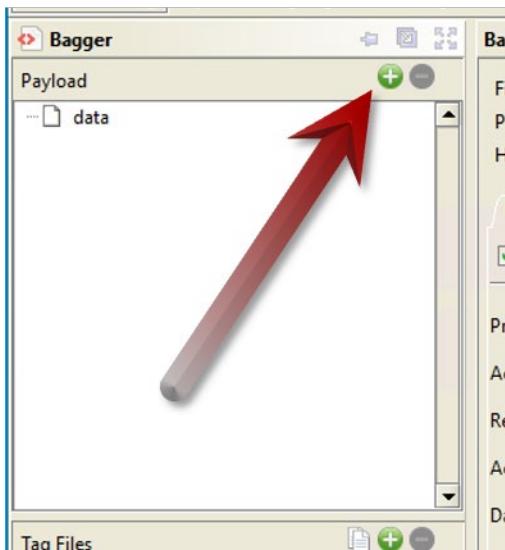
- We now have the skeleton of the bag
 - a Payload area with a blank data folder
 - a Tag Files area with empty bag-info.txt and bagit.txt files
 - a Bag-Info area where you will enter the minimum metadata elements for the bag



Bagger Figure 11: Bag skeleton framework

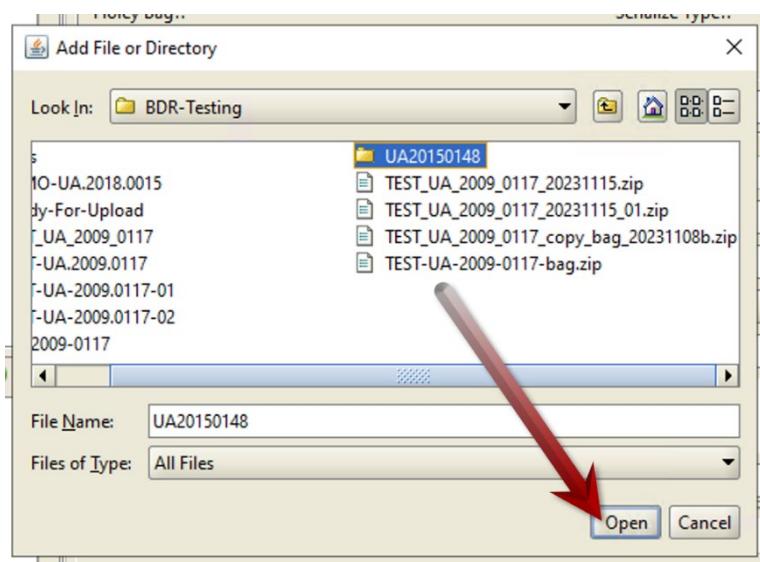
Adding a Payload

- In the upper right corner of the Payload section, click the green plus (+) button to add the folder(s) file(s) you wish to bag.



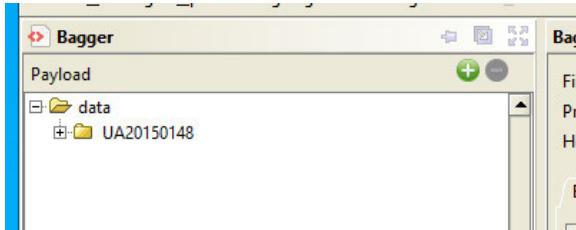
Bagger Figure 12: Adding folders and files to Payload

- This will open a dialog window, allowing you to navigate to the folder of material you want to add. Once you find the folder of content you want to bag, select Open.



Bagger Figure 13: Adding folders and files to bag

- We now see that the data folder in the Payload section is now populated with the folder you selected



Bagger Figure 14: Bagger's Payload section with data folder populated with folders and files

Adding Bag-Info

- Next you will need to complete the metadata for the bag in the Bag-Info section
 - Default fields
 - Profile Name
 - Source-organization
 - Required fields will be indicated by a red “R”
 - Other common fields will include
 - Accession Number (or collection identifier)
 - Access Restrictions; chosen from a pull-down menu
 - Open/Public
 - Restricted/Confidential
 - Not Yet Known
 - Contains Some Confidential Records



Bagger Figure 15: Access Restrictions pull-down menu

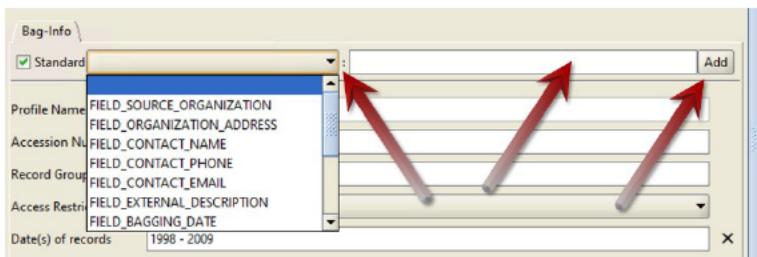
- Bag creator
- Other fields as required
 - e.g. Record Group for University Archives
- Other optional fields
 - e.g. Date(s) of records

The screenshot shows the 'Bag-Info' configuration window with the 'Standard' profile selected. The fields filled in are:

- Profile Name: OSU-Archive
- Accession Number: UA-2015.0148
- Record Group: 5.a
- Access Restrictions: Open/Public
- Date(s) of records: 2013 - 2015
- Source-organization: The Ohio State University, University Archives
- Bag Creator: Danny the DP Archivist

Bagger Figure 16: Bag-Info metadata fields completed

- While we do not recommend it, additional metadata fields could be added at this point by clicking on the “Standard” pull-down menu, selecting a field, adding a value in the text box and finally clicking the “Add” button.

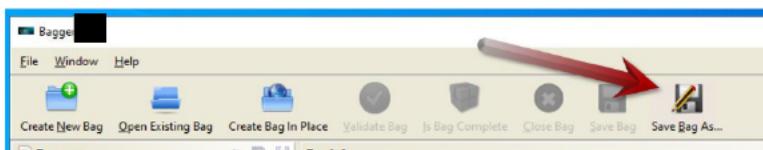


Bagger Figure 17: Adding a metadata field to the bag profile

Saving Bag

Having completed the input for the Bag-Info, it is time to save the bag.

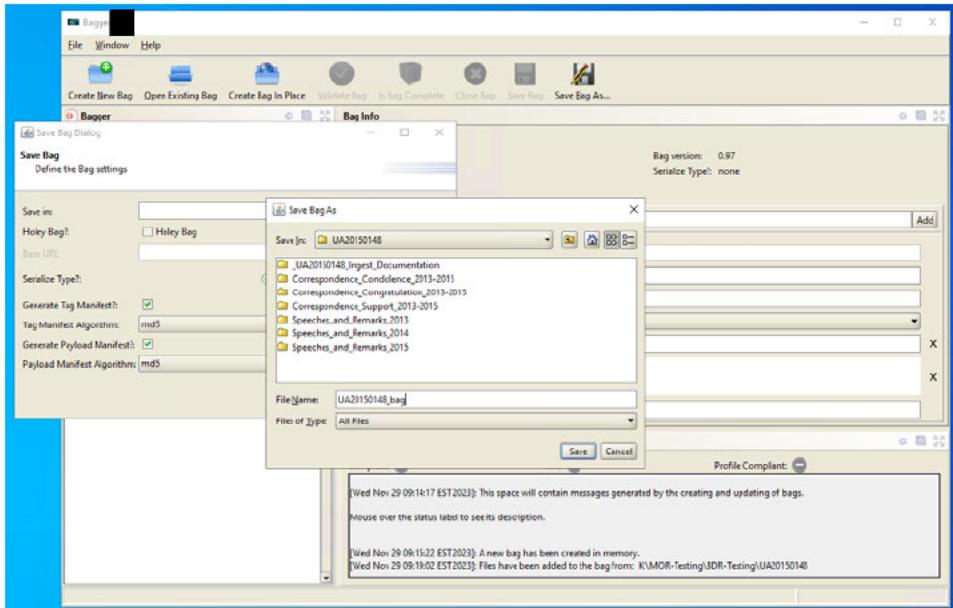
- Click “Save Bag As...” at the top right of the toolbar.



Bagger Figure 18: Selecting "Save Bag As..."

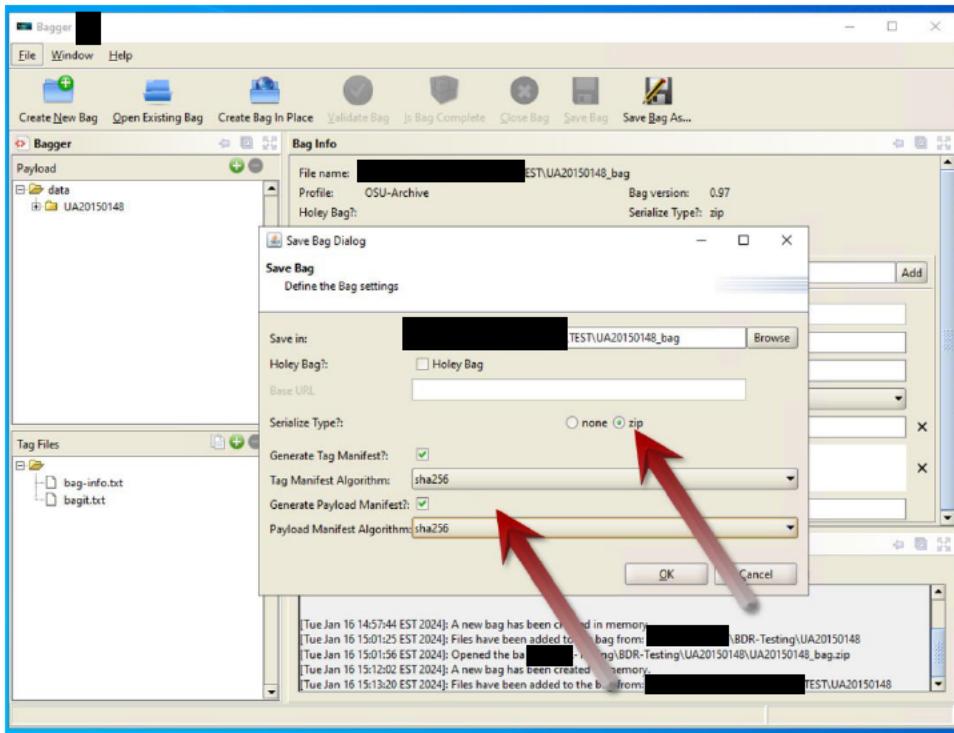
- A “Save Bag” dialog window will now be open. Click the Browse button and navigate to the root directory of the files being bagged, name the bag with the accession number with a “_bag” suffix (e.g. UA20150148_bag), and select “Save.” **NOTE: the bag’s file name SHOULD NOT contain a period (.)**; if it does

it will throw off the zipped bag results. Replace any periods with an underscore (_).



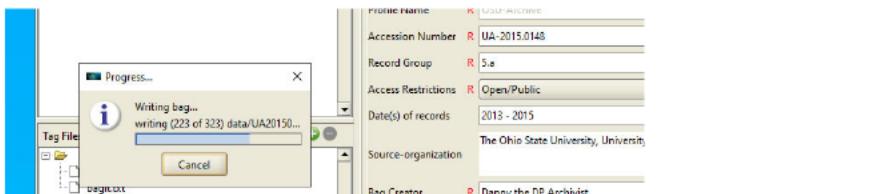
Bagger Figure 19a: "Save Bag As..." features and selections

- Leave the “Holey Bag” box unchecked.
- “Serialize Type” defaults to “none”; click the “zip” radio button.
- Make sure the “Generate Tag Manifest?” and “Generate Payload Manifest?” boxes are checked. This will create tagmanifest and manifest .txt files.
- The “Tag Manifest Algorithm” and “Payload Manifest Algorithms” likely will be defaulted to md5, but we will use sha256.



Bagger Figure 19b: "Save Bag As..." features and selections

- Select OK. Bagger will now create the bag



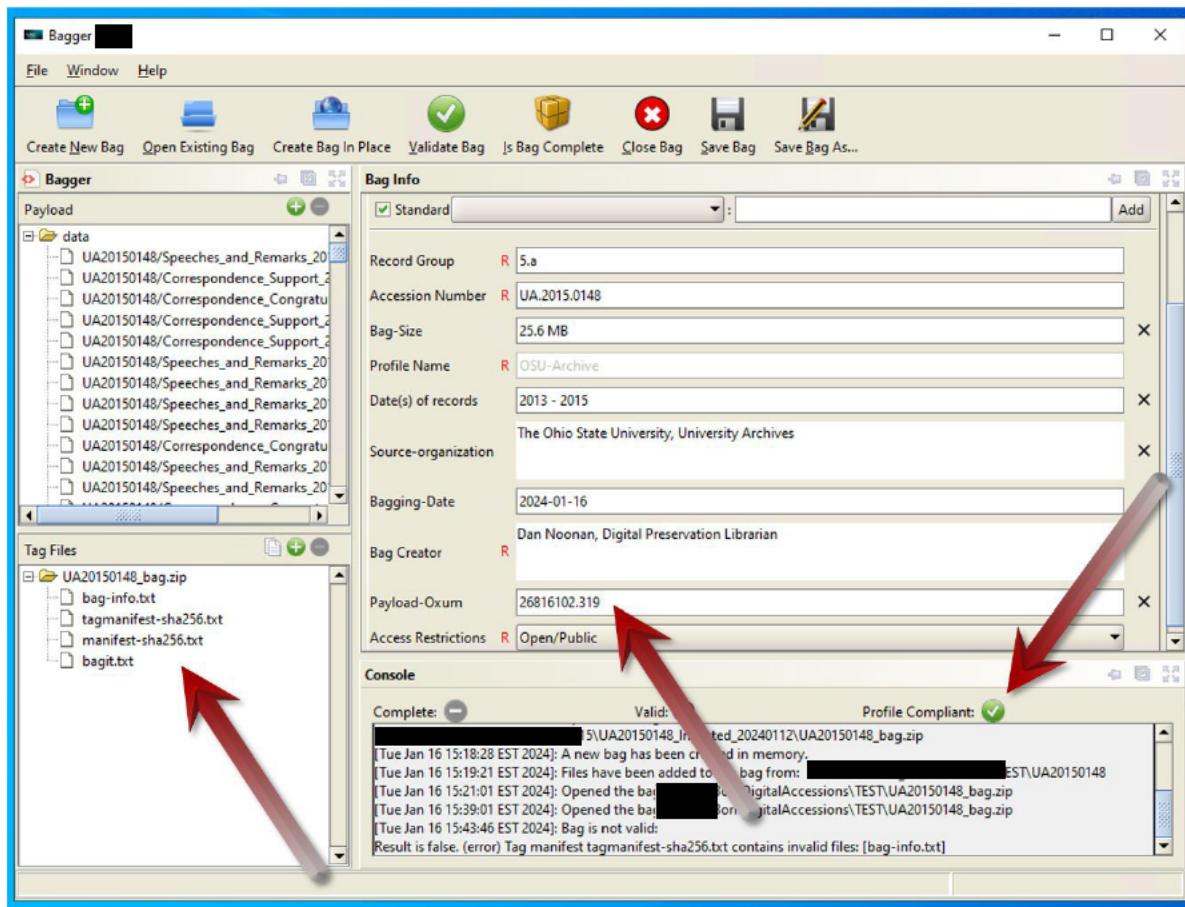
Bagger Figure 200: Writing bag...

Completed Bag

Once Bagger has completed making the bag you will see some changes to the interface.

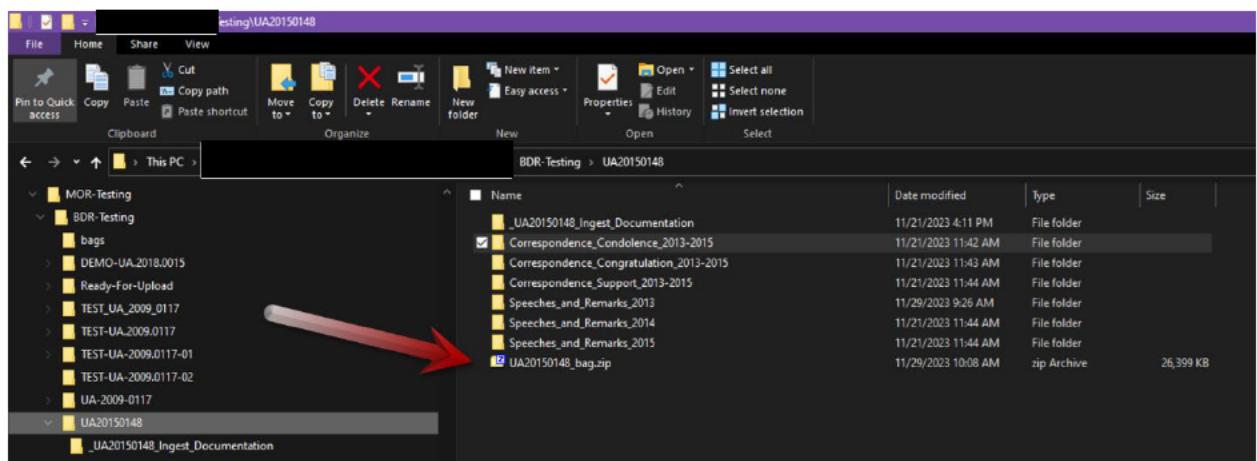
- The Tag Files section now has the addition of the tagmanifest-sha256.txt and the manifest-sha256.txt files
- The Profile Compliant note in the Console section should now have a green checkmark
- The Bag-Info section will have refreshed and re-arranged itself to now include the
 - Bag-Size - in this example it is 25.6 MB

- Payload-Oxum. The Payload-Oxum is an interesting number; it is the "octetstream sum" of the payload, namely, a two-part number, OctetCount.StreamCount, where OctetCount is the total number of octets (8-bit bytes) across all payload file content and StreamCount is the total number of payload files. So, in this example the Payload-Oxum = 26816102.319; we have 26,816,102 bits of file size and 319 files.
 $26,816,102 \text{ bits} = 26,187.60 \text{ bytes} = 25.57 \text{ MB}$
- The Payload-Oxum is written to bag-info.txt file.
- We will be including this information in the [Local Administrative Dashboard](#)



Bagger Figure 21: Bagger interface post-bag creation

- The bagged content can now be found in the root folder, and is ready for transfer and ingest into the Gray Repo



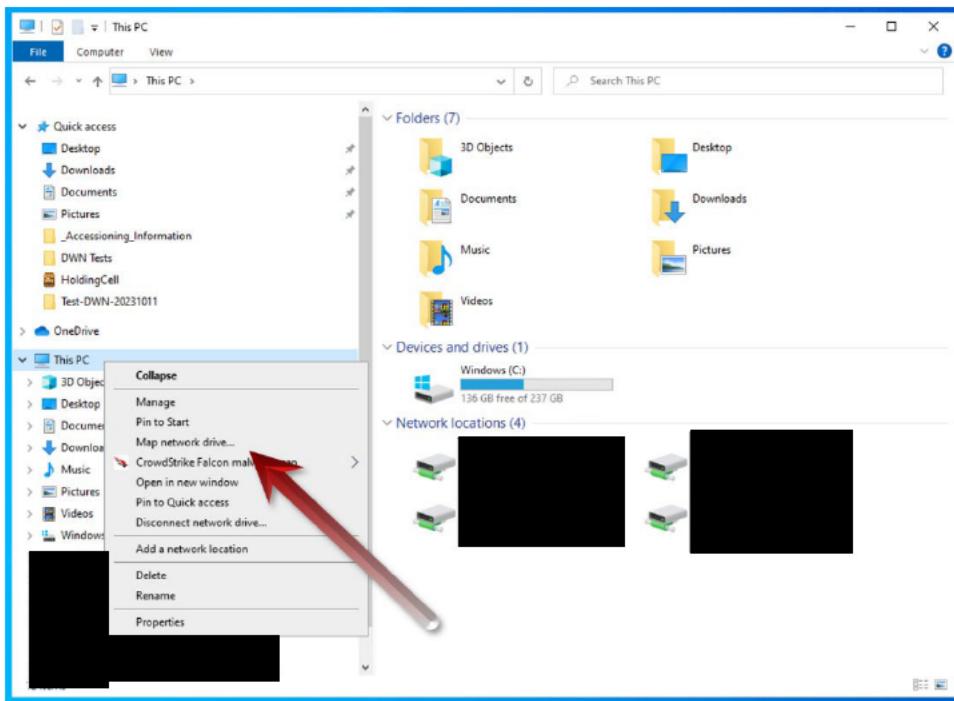
Bagger Figure 22: Bagged content in the root folder ready for transfer and ingest into the Gray Repo

Ingest

Mapping [REDACTED] bucket

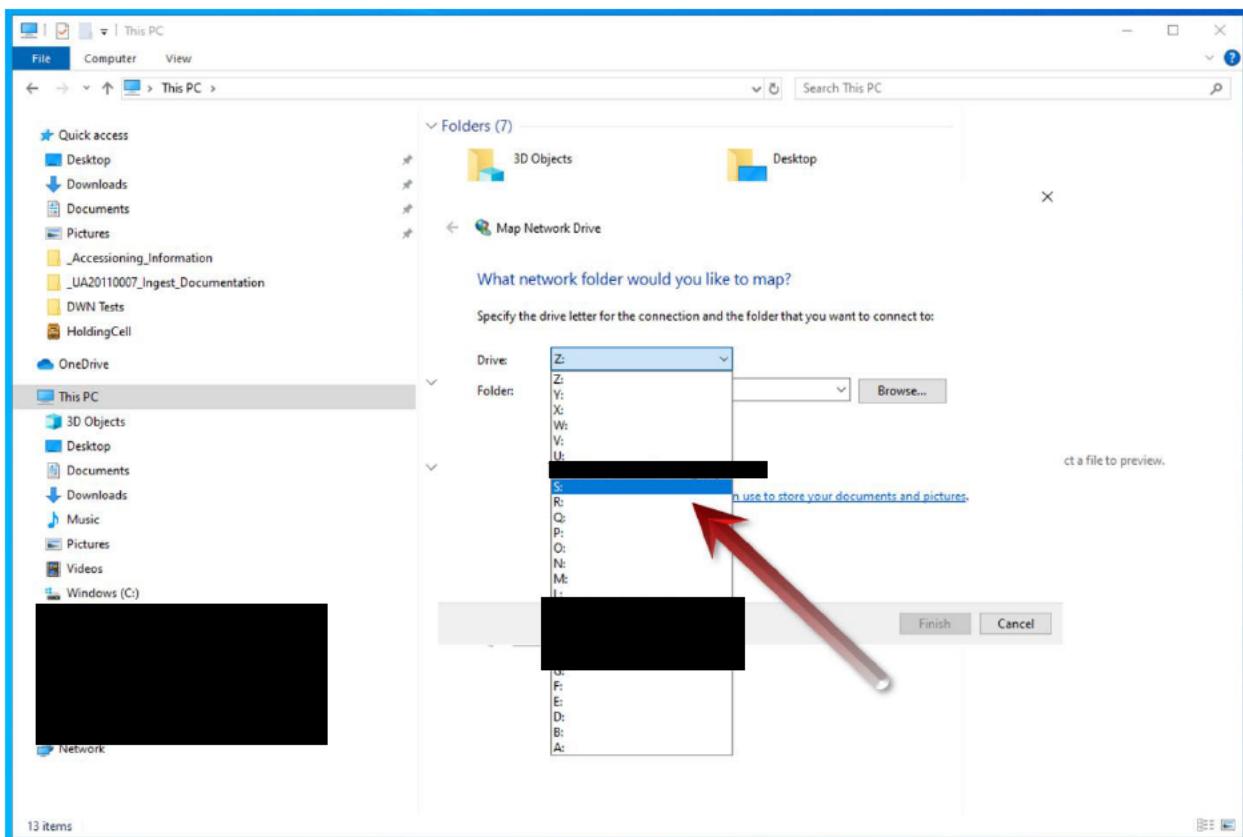
As previously mentioned, the ingest point for the Gray Repo will be what is called an AWS [REDACTED] Bucket. While it is not on the University's network, we can still map it as if it is.

- Connect to the [Remote Desktop](#)
- You have to be logged in to [REDACTED] VPN on the Remote Desktop (even if you connected to [REDACTED] to get to the Remote Desktop) to access, map and use this network drive.
- In Windows Explorer, navigate to “This PC”
- Right click on This PC
- Select and click on “Map network drive ...”



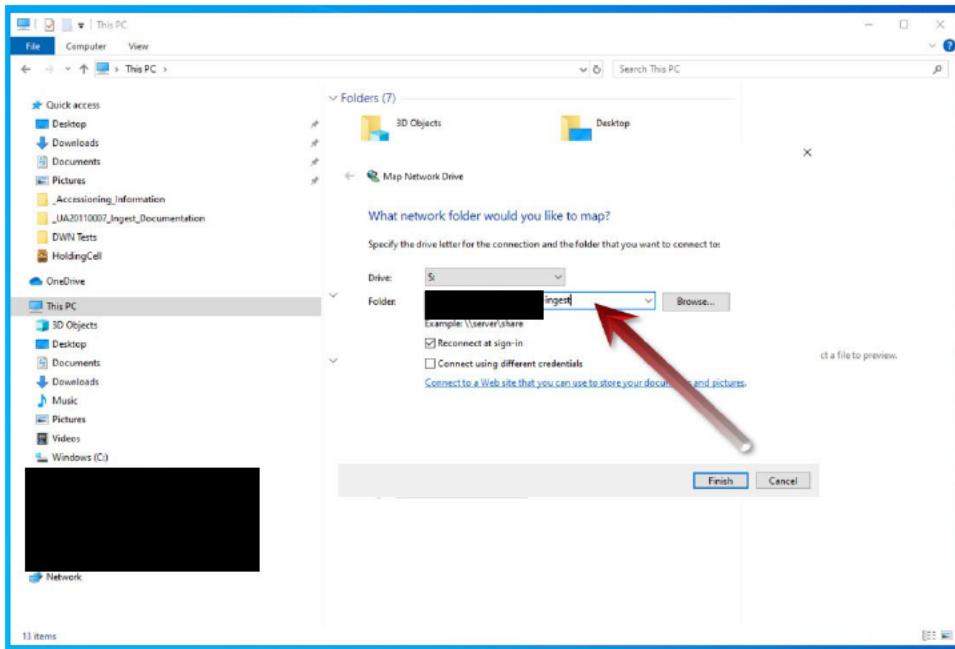
[REDACTED] Bucket Mapping Figure 1: Selecting "Map network drive..."

- A dialog box will open allowing you to choose a drive letter and destination



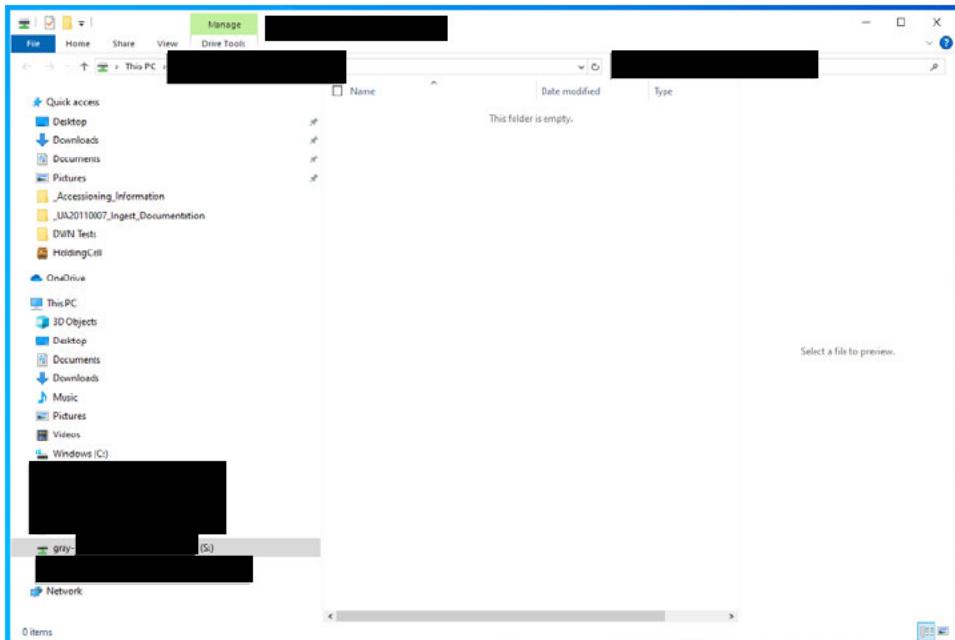
■ Bucket Mapping Figure 2: Map Network Dive dialog box

- Choose an unused drive letter from the pull-down menu; we suggest using S to coincide with █ for the production environment, and T if you are connecting to the training sandbox version of the Gray Repo.
- Enter the network address:
 - Production Environment: █
 - Training Environment: █



▀ Bucket Mapping Figure 3: Map Network Drive with desired information entered

- If not already checked, check the “Reconnect at sign-in” box
- Click Finish
- You should now have a mapped S-drive to the █ bucket

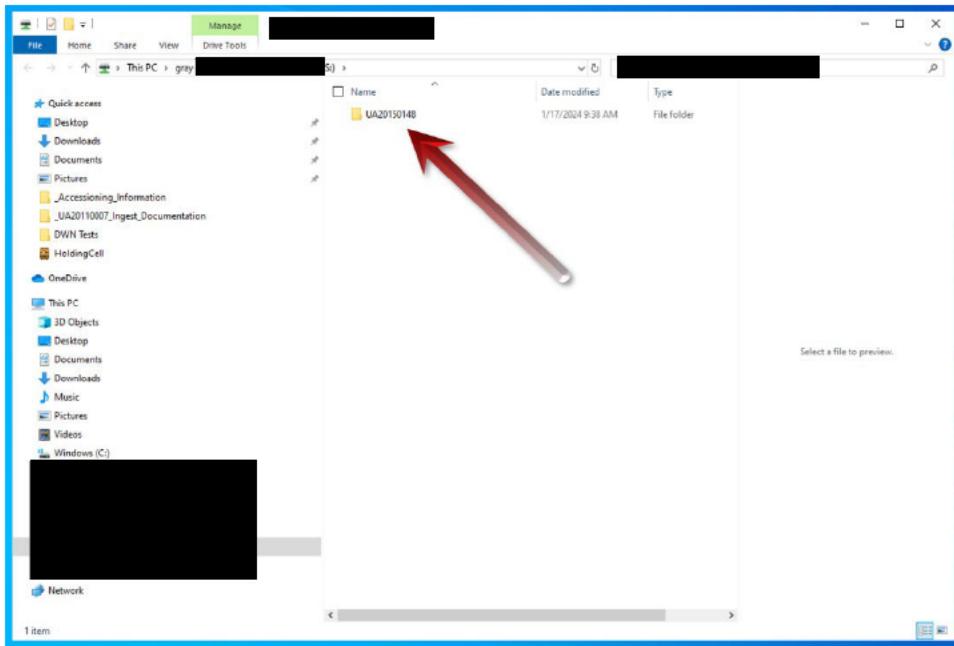


▀ Bucket Mapping Figure 4: Mapped S-drive to █ bucket

Ingest

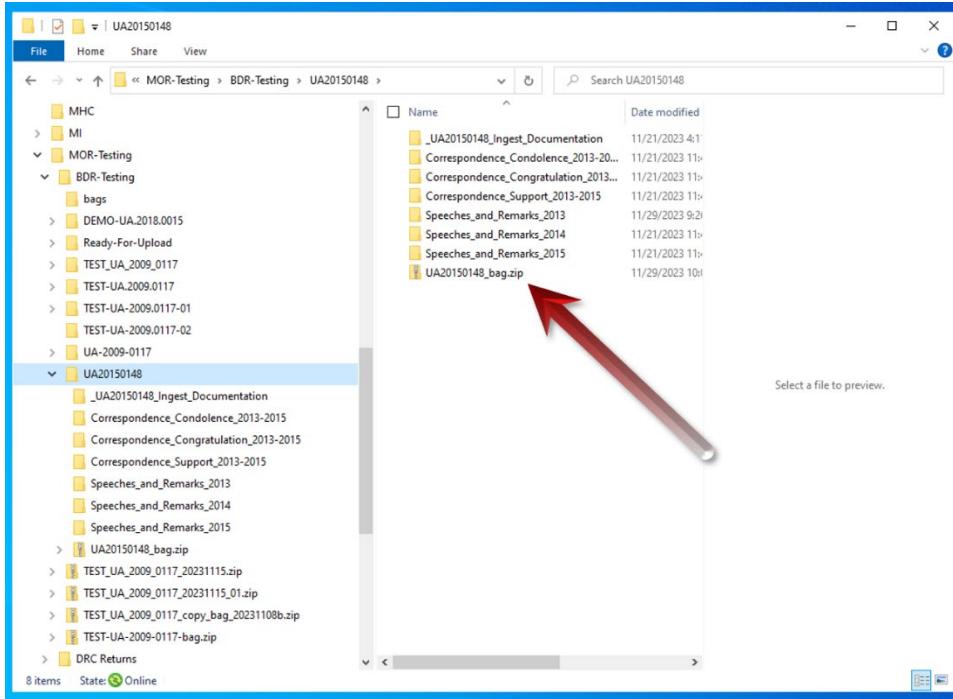
The actual act of ingesting a bag into our Gray Repo is probably the simplest step within this workflow; it is essentially a copy and paste of the zipped bag from its root directory to the [REDACTED] bucket. That act of placing a new zip file in the [REDACTED] bucket triggers the ingest process into the Gray Repo, the Fedora platform environment we have established on AWS. When the ingest process in Fedora has completed it will generate a brief text file back to the ingest directory in the [REDACTED] bucket. This text file will be both named with the Fedoraid and contain it; the Fedoraid allows for future retrieval of the bag.

- Navigate to the [REDACTED] bucket (S-drive or the letter you have named it)
- Create a temporary folder for the ingest of your bag. You should name the temporary folder the same as your initial accession folder.



Ingest Figure 1: Create temporary ingest folder

- Next navigate to your accession's root folder and locate the bag file



Ingest Figure 2: Identify and Copy bag

- Copy that file.
- Navigate back to the temporary folder you created in the [] bucket and paste the bag.
- Next all the “magic” happens in the background.
- You can check to see if the ingest is progressing at:
 - Production Environment: [].
 - Training Environment: contact Digital Preservation Department



Fedor 6 Inventory

File	Fedor ID	Checksum	Ingested	Ingesting...
Test-DWN-20231011/TEST-UA-2009.0117-bag.zip	d2f40203-d329-4e77-ad2e-4fef9be9593e	✓	✓	
Test-DWN-20231011/TEST-UA-2009.0117-bag.zip	3b210323-cff60-4175-94cf-f00fb24a8ac	✓	✓	
Test-DWN-20231011/TEST-UA-2009-0117-bag.zip	652625ed-29a7-498e-93d-578fde4ee30a	✓	✓	
TEST-UA-2009.0117/TEST-UA-2009-0117-bag.zip	1ea4c819-b8c1-4e1d-8664-811103d957b6	✓	✓	
UA_2009_0117/UA-2009-0117-bag.zip	6ec72223-138a-433b-9d3e-579d118cb51f	✓	✓	
SPEC-12235/squirrel.zip	6a9b550c-04eb-4c58-8fc7-e9c1b2948a4	✓	✓	
forTESTINGonly/TEST_UA_2009_0117_copy.bag.20231108azip	8cee8282-356c-4e64-a061-00230f55fa5c	✓	✓	
forTESTINGonly/TEST_UA_2009_0117_copy.bag.20231108bzip	042ea28c-ccaa-4ffd-98d7-897421ba3df5	✓	✓	
SPEC-122356/squirrel.zip	53c89943-5e5a-4f6d-9916-42299d8ed2f5	✓	✓	
dwnTEST/TEST_UA_2009_0117_20231115.zip	70fae428-c945-4cd0-be52-644c34dbaab0	✓	✓	
dwnTEST/TEST_UA_2009_0117_20231115_01.zip	9109259cfea8-4043-88ce-0fec30baba5	✓	✓	
UA-2009-0117/UA-2009-0117-bag.zip	c20322f9-6e3f-4974-a624-fc20146a402b	✓	✓	
UA-2009-0117-TEST/UA-2009-0117-bag.zip	a392fc35-b083-4e62-b0ca-86bdd2661bdb	✓	✓	
UA20150148/UA20150148_bag.zip	b1ff2af5-8174-48b7-ba28-08b85c0341e3	✓	✓	
dwnTEST/UA20150148_bag.zip	f8c09d1c-99ec-417c-9319-6b3bf1ac125d	✓	✓	
UA20150148_TEST/UA20150148_bag.zip	-	-	-	⌚

Ingest Figure 3: Fedora ingest in progress



dwnTEST/TEST_UA_2009_0117_20231115.zip	70fae428-c945-4cd0-be52-644c34dbaab0	✓	✓
dwnTEST/TEST_UA_2009_0117_20231115_01.zip	9109259cfea8-4043-88ce-0fec30baba5	✓	✓
UA-2009-0117/UA-2009-0117-bag.zip	c20322f9-6e3f-4974-a624-fc20146a402b	✓	✓
UA-2009-0117-TEST/UA-2009-0117-bag.zip	a392fc35-b083-4e62-b0ca-86bdd2661bdb	✓	✓
UA20150148/UA20150148_bag.zip	b1ff2af5-8174-48b7-ba28-08b85c0341e3	✓	✓
dwnTEST/UA20150148_bag.zip	f8c09d1c-99ec-417c-9319-6b3bf1ac125d	✓	✓
UA20150148_TEST/UA20150148_bag.zip	eace3ae3-478d-4766-8c05-24bfaadc269	✓	✓

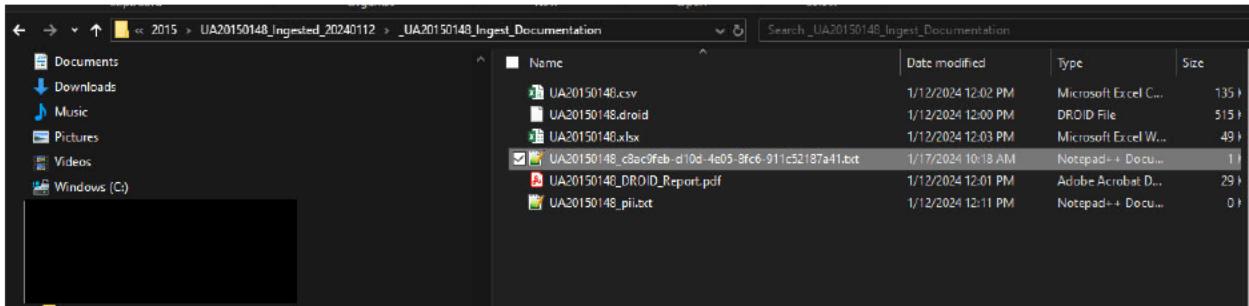
Ingest Figure 4: Fedora ingest complete

- Depending upon the size of the bag, the ingest time may be lengthy. You will know the process is complete when the FedorID text file appears in your temporary folder. For small ingest the FedorID should show up within 5 to 10 minutes.

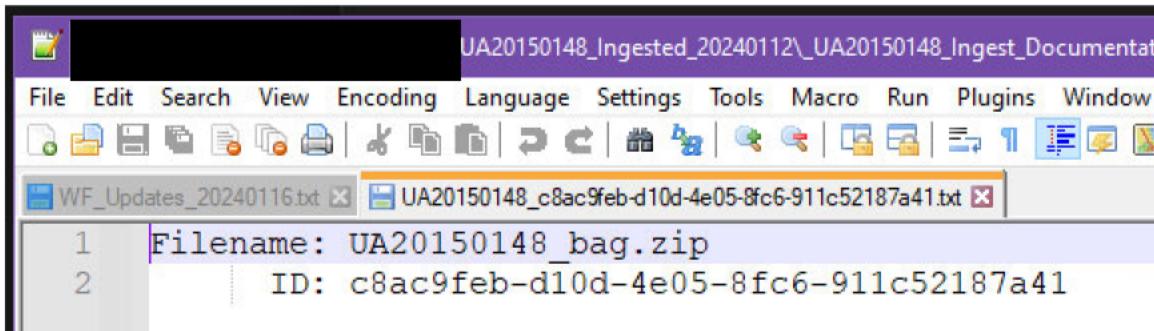
FedorID

The FedorID will arrive back in our temporary ingest folder as a text file; it is machine-generated named, with a string of gobbledegook, alpha-numeric characters (kind of like a checksum; e.g. c8ac9feb-d10d-4e05-8fc6-911c52187a41.txt). You should rename it

to include the Accession ID as prescribed in the file naming section above, “AccessionID_[FederalID].txt” (e.g. UA20150148_c8ac9feb-d10d-4e05-8fc6-911c52187a41.txt), and copy it to the Ingest Documentation folder



Ingest Figure 5: FederalID text file renamed with AccessionID appended



Ingest Figure 6: The FederalID text file contents

You can open the FederalID text file in a text editor such as Notepad, Wordpad or Notepad++. It is a simple file that contains the filename of the payload and the FederalID. The FederalID will be recorded in the [Local Administrative Dashboard](#), as well as in Archivist Toolkit or PastPerfect.

Finding Aid Linkage

How will our patrons/researchers/users know that we have these records preserved and available for their research? First, as per standard University Libraries accessioning procedures an accession record will be created within Archivist Toolkit or PastPerfect depending upon the collecting unit. In reviewing a finding aid the user will know that these records exist. The curatorial staff will know where to retrieve them as the FedorID will be included in the record.

Archivist Toolkit

The Archival Technical Services Department has created a field within Archivist Toolkit to record the FedorID. It is located in the “User defined fields” tab.



Finding Aid Figure 1: FedorID field in Archivists Toolkit

PastPerfect

 NOTE: This section UNDER CONSTRUCTION.
The Billy Ireland Cartoon Library and Museum utilizes PastPerfect for its accessioning records. The Digital Preservation Department will develop a procedure in consultation with their staff.

Retrieval

When a user/patron/researcher indicates to the curatorial staff that they would like to view records stored in the Gray Repo, the curatorial staff can refer to the finding aid or the Gray Repo Admin Console to identify the appropriate FedorID to retrieve. Further, they should determine if there are any PII restrictions that need to be dealt with prior to providing copies to the user/patron/researcher.

Accessing the Gray Repo

The Gray Digital Preservation Repository is at [REDACTED] and requires login with Ohio State name.# credentials and DUO authentication. Further, access is restricted to appropriate curatorial, digital preservation and administrative Technology & Digital Programs staff. The interface is rudimentary, simply a listing of the bags and their FedorIDs, along with ingested date and indicators for checksum generation and ingestion completeness.

File	Fedor ID	Ingested Date	Checksum	Ingested	Ingesting...
UA20110007/UA20110007_bag.zip	d34fe7a7-b936-40fe-aef9-d43b8fd46af4	January 12 2024	✓	✓	
UA20110006_bag.zip	e556fcf5-d9cc-4176-81fa-0e359e2b6871	January 12 2024	✓	✓	
UA20150148/UA20150148_bag.zip	c8ac9feb-d10d-4e05-8fc6-911c52107a41	January 12 2024	✓	✓	
UA20090117/UA20090117_bag.zip	81c8a219-1593-4a81-849a-1aa297b4e299	January 12 2024	✓	✓	

Retrieval Figure 1: Gray Repo Retrieval Interface

- Log-in at the aforementioned URL
- As there is no traditional search functionality, you will need to do a “Find” (Chrome) or “Find in page” (Firefox) using Ctrl+F
- Enter FedorID in the search box and the entry will be highlighted on the page

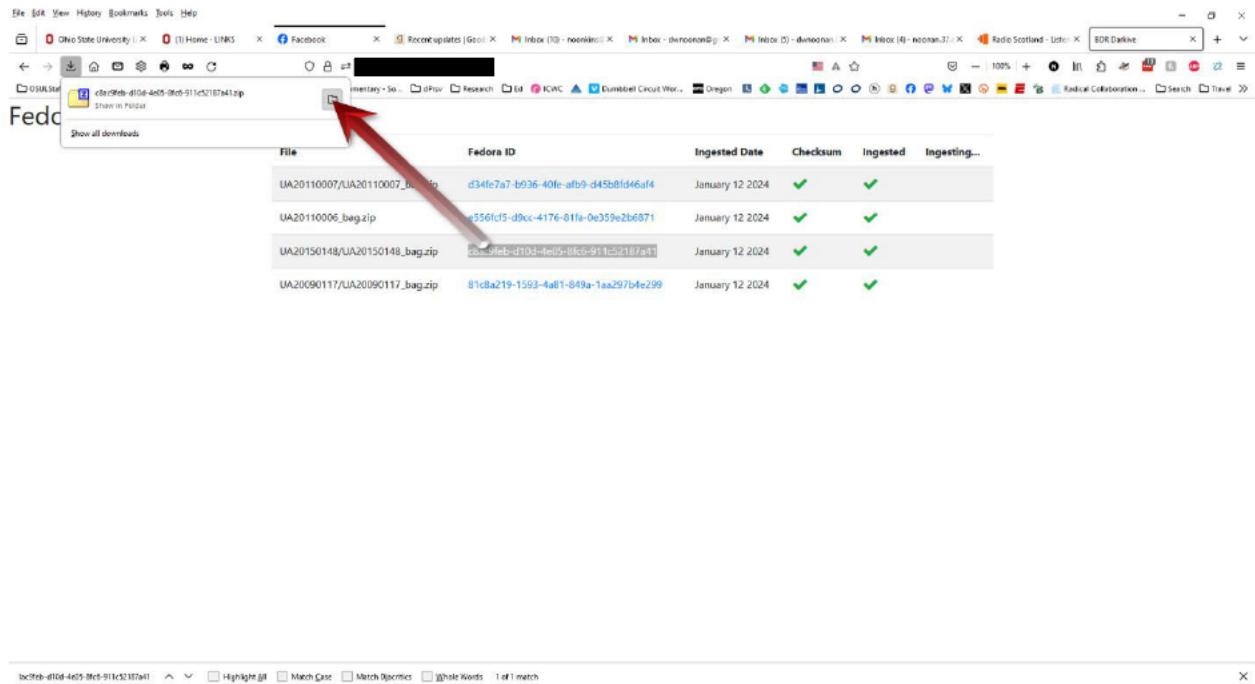
File	Fedor ID	Ingested Date	Checksum	Ingested	Ingesting...
UA20110007/UA20110007_bag.zip	d34fe7a7-b936-40fe-aef9-d43b8fd46af4	January 12 2024	✓	✓	
UA20110006_bag.zip	e556fcf5-d9cc-4176-81fa-0e359e2b6871	January 12 2024	✓	✓	
UA20150148/UA20150148_bag.zip	c8ac9feb-d10d-4e05-8fc6-911c52107a41	January 12 2024	✓	✓	
UA20090117/UA20090117_bag.zip	81c8a219-1593-4a81-849a-1aa297b4e299	January 12 2024	✓	✓	



Retrieval Figure 2:: FedorID search and result using "Find in page"

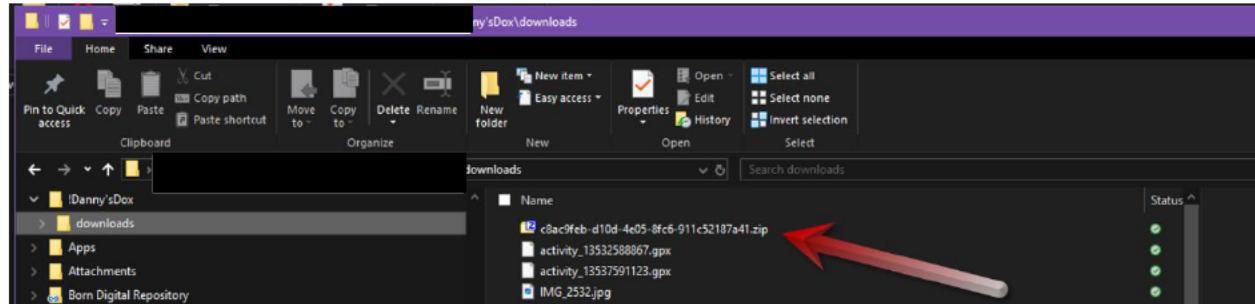
- Click on the highlighted FedorID and the bag will be downloaded

- When the download is complete, open the download folder



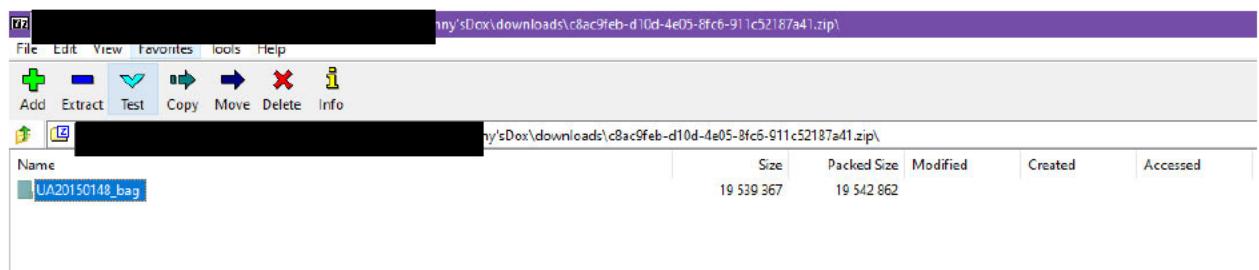
Retrieval Figure 3: Open download folder to access Bag

- There will be a zip file whose file name is FedorID

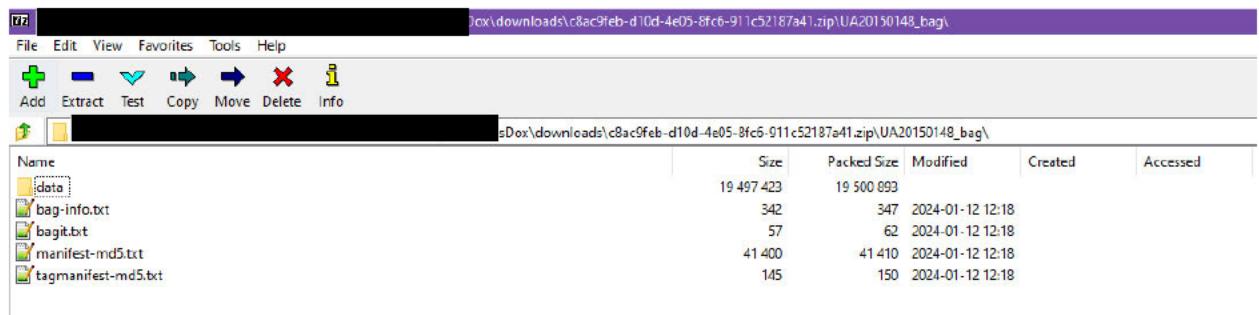


Retrieval Figure 4: Downloads folder with retrieved Bag

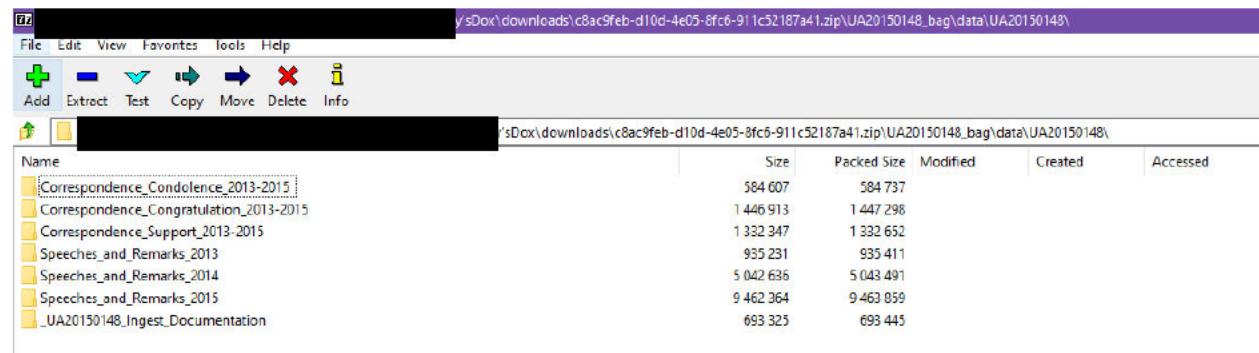
- The file will need to be unzipped (preferably using 7zip, as native Windows unzipping may provide undesirable results for files created in a Mac environment) and the contents extracted to be shared with the patron/user/researcher



Retrieval Figure 5: The Bag within downloaded zip file



Retrieval Figure 6: Bag contents



Retrieval Figure 7: Bagged data

Local Administrative Dashboard

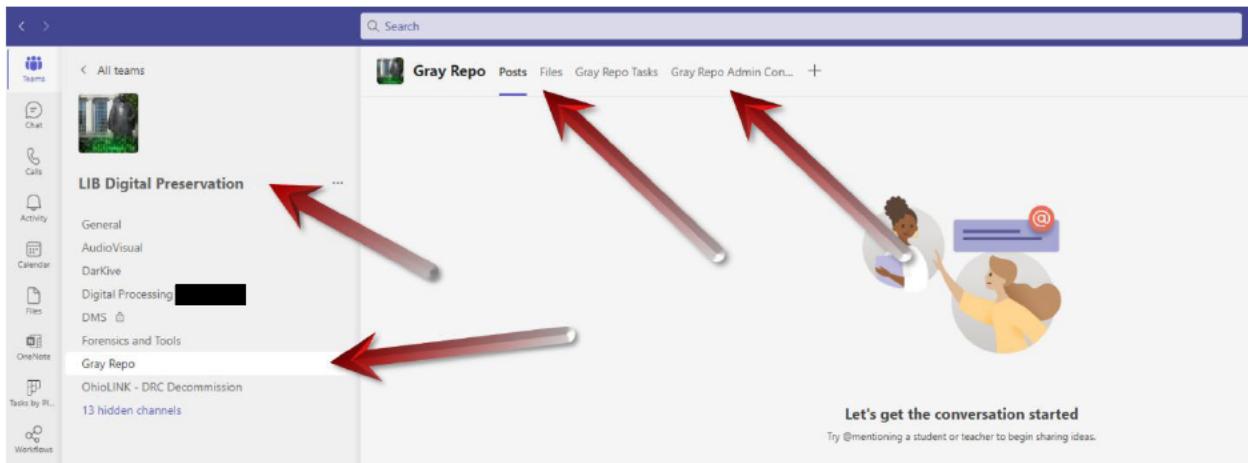
There are two major differences between the former so-called Dark Archive and the Gray Digital Preservation Repository:

- The Gray Repo is built intentionally for digital preservation, not utilized for such purposes as an afterthought.
- We have intellectual and administrative control over the objects ingested into the Gray Repo.

In regards to the latter point, intellectual control is provided through the FedorID linkage to the finding aid, while administrative control is provided through the Local Administrative Dashboard. The initial iteration of this administrative dashboard is a combination of tools within Microsoft Teams

- Teams: [LIB Digital Preservation](#)
 - Channel: [Gray Repo](#)
 - Files: [DashboardFiles](#)
 - Lists: [Gray Repo Admin Console](#)

As appropriate people will be added to the Team/Channel. Contact Digital Preservation if you have issues with access or use of the tools.



Local Administrative Dashboard Figure 1: Local Administrative Dashboard components

Files Maintained Locally

As noted in the [Ingest Documentation Folder](#) section above, the files located in the _AccessionID_Ingest_Documentation folder will be copied and maintained in the Local Administrative Dashboard:

- AccessionID.droid
 - Maintained for digital preservation administrative purposes
- AccessionID_debug_file.txt (if necessary)
 - Maintained for processing provenance purposes
- AccessionID_DROID_Report.pdf
 - Maintained for digital preservation administrative purposes
- AccessionID_DROID_Report.xlsx or AccessionID_DROID_Report.csv
 - Sharable with potential patrons/users/researchers as a first line of research before files are downloaded
- AccessionID_[FedoralID].txt
 - Maintained for digital preservation administrative and processing provenance purposes
- AccessionID_pii.txt
 - If personally identifiable information had been discovered, this allows curatorial staff to pinpoint what may need to be redacted
- AccessionID_ReadMe_YYYYMMDD.txt
 - Maintained for digital preservation administrative and processing provenance purposes

These files will be held within the _DashboardFiles folder within the LIB Digital Preservation>Gray Repo Teams Channel. This folder has been sub-divided with folders for the various collecting units within University Libraries. If an area is not represented please contact Digital Preservation to have it added.

Name	Modified	Modified By
_DashboardFiles	Yesterday at 8:53 AM	Cashen, Kaleb
Cornell Workflow Example	February 16	Noonan, Dan
files to Migrate	November 15, 2022	Noonan, Dan
OTP	May 17	Noonan, Dan
Resources	October 12	Noonan, Dan
Service description	October 12	Noonan, Dan
Wiki-Images	June 15	Noonan, Dan
Workflow	October 12	Noonan, Dan
RetrievalError_20231127a.PNG	November 27	Noonan, Dan
RetrievalError_20231127b.PNG	November 27	Noonan, Dan

Local Administrative Dashboard Figure 2: *_DashboardFiles* folder in Gray Repo channel

Name	Modified	Modified By
ARV	November 17	Noonan, Dan
BPA	November 17	Noonan, Dan
CGA	November 17	Noonan, Dan
GEN	November 17	Noonan, Dan
HIL	November 17	Noonan, Dan
KB	November 17	Noonan, Dan
MUS	November 17	Noonan, Dan
OPPA	November 17	Noonan, Dan
RAR	November 17	Noonan, Dan
TRU	November 17	Noonan, Dan

Local Administrative Dashboard Figure 3: University Libraries unit-specific folder within *_DashboardFiles*

It is within the collecting unit's folder that the ingest documentation will be stored. The most efficient way of transferring this documentation is to:

- copy the *_AccessionID_Ingest_Documentation* to the unit's folder
 - e.g. *_UA20150148_Ingest_Documentation*
- rename it to the AccessionID
 - e.g. *UA_2015_0148*

A screenshot of a Microsoft Teams interface. On the left, there's a sidebar with icons for Chat, Calls, Activity, Calendar, File, OneNote, and Tasks by PL. The main area shows a team named 'LIB Digital Preservation'. Under this team, there are several sections: General, AudioVisual, DarKive, Digital Processing (with a redacted icon), DMS, Forensics and Tools, and Gray Repo. The 'Gray Repo' section is expanded, showing a file list with two entries: '_UA_2015_0148_Ingest_Documentation' and 'UA_2099_0117'. Both files were modified by 'Noonan, Dan'. The file names are preceded by small yellow folder icons.

Local Administrative Dashboard Figure 4: _AccessionID_Ingest_Documentation copied to unit's folder

A screenshot of a Microsoft Teams interface, similar to Figure 4 but with a different file structure. The 'Digital Processing' category in the sidebar has a redacted icon. The 'Gray Repo' section shows a file list with two entries: 'UA_2015_0148' and 'UA_2099_0117'. Both files were modified by 'Noonan, Dan'. The file names are preceded by small yellow folder icons.

Local Administrative Dashboard Figure 5: _AccessionID_Ingest_Documentation renamed to AccessionID

A screenshot of a Microsoft Teams interface, similar to Figures 4 and 5. The 'Digital Processing' category in the sidebar has a redacted icon. The 'Gray Repo' section shows a file list with several items: 'UA20150148.droid', 'UA20150148_c8adfeb-f10d-4e05-9c6-911c52187a41.txt', 'UA20150148_DRDID_Report.csv', 'UA20150148_DRDID_Report.pdf', 'UA20150148_DRDID_Report.xlsx', and 'UA20150148_pii.txt'. All files were modified by 'Noonan, Dan'. The file names are preceded by small icons representing their type (e.g., document, CSV, PDF).

Local Administrative Dashboard Figure 6: Files maintained in AccessionID folder

Data Maintained in Gray Repo Admin Console

We will be maintaining a limited amount of administrative data in a Microsoft List to assist both the curatorial and digital preservation staff. There are eleven (11) required data elements, along with an optional Notes field:

- Required
 - Bag: The bag file name sans the “.zip”.
 - Unit: The responsible collecting unit/area within University Libraries. It is a controlled vocabulary, radio button choice
 - Collection: Name of the collection the accession is part of.
 - AccessionID
 - Fedoraid
 - PII Restrictions: A radio button selection of “Yes” or “None Found”
 - #files: the number of files derived from the [Payload-Oxum](#)
 - Bits: the number of bits derived from the [Payload-Oxum](#)
 - Bytes, KBs, MBs, GBs and TBs are calculated and stored from this number
 - Ingest Documentation: a link to the ingestion documentation folder; for simplicity the alternative text should be “Files”
 - Ingest Date
 - Ingested by
- Optional:
 - Notes: While this is a free-form multi-line text field, for consistency use the following formula for entering notes.
 - YYYYMMDD: [initials of commenter; e.g. DN for Dan Noonan] comments/notes (e.g. 20231201: [DN] This is the Notes field.)

There are two ways to interact with Microsoft Lists:

- within the Teams channel (either in the desktop app or online) in the navigation bar “Gray Repo Admin Console”

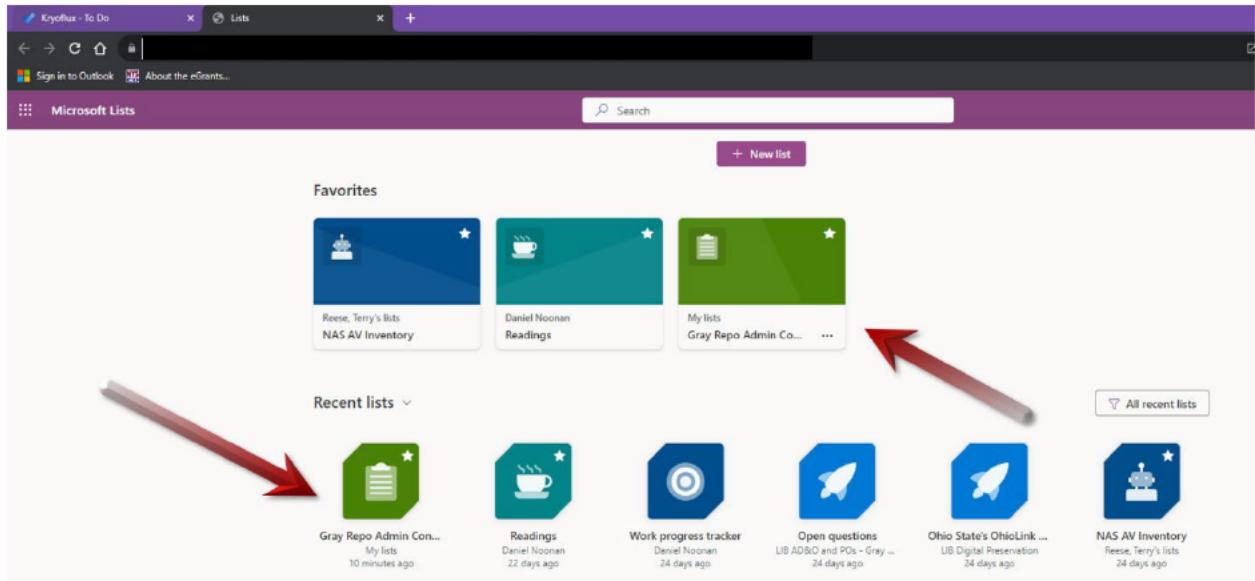
Bag	Unit	Collection	AccessionID	FederalID	PI Restrictions	Bits	#Files	Bytes
UAJ0090117_bag	[ARV]	Career Connection	UA.2009.0117	81c8a219-1593-4861-849e-1a297e4e299	None Found	40,609,399	168	39,658
UAJ0110005_bag	[ARV]	Faculty Council	UA.0211.0005	e156fcf5-49cc-4176-818a-04359eb6871	None Found	2,168,107	7	2,117
UAJ0110007_bag	[ARV]	University Senate	UA.2011.0007	c14fe7a7-4b34-4ff6-a709-d45bd645a44	None Found	6292,390	10	6,145
UAJ0151048_bag	[ARV]	Office of the Provost	UA.2015.0148	cfa9feb-d10d-4e05-8fc6-911c52107a41	None Found	26,821,745	320	16,193

Local Administrative Dashboard Figure 7: Gray Repo Admin Console List within Teams desktop app

- in the List app online, found in the Microsoft 365 App Launcher; the Gray Repo Admin Console List can be added to your “Favorite” Lists

The Microsoft 365 App Launcher interface shows the 'My Day - To Do' task card on the left and the 'Gray Repo Dashboard' card on the right. The 'Apps' section lists various Microsoft services. A red arrow points to the 'Lists' icon, which is represented by a yellow square with a blue grid pattern.

Local Administrative Dashboard Figure 8: Locating Lists app in Microsoft 365 App Launcher



Local Administrative Dashboard Figure 9: Microsoft 365 Lists App

Bag	Unit	Collection	Accession#	FederalID	PI Restrict.	Bits	Files	Bytes	KBs	MBs	GBs	TBs
UA20090117_bag	ARV	Career Connection	UA2009.0117	815a131b-1934-4ed1-80f9-1aa27d4e299	None Found	40,601,399	158	39,651	39	0.04	0.000	0.0
UA20110006_bag	ARV	Faculty Council	UA0211.006	e556ff5d-0910-4176-811a-0a150a2a671	None Found	2,166,107	7	2,117	2	0.00	0.000	0.0
UA20110007_bag	ARV	University Senate	UA2011.007	d349c7a7-b916-40fe-ab50-d45d8966da4	None Found	6,292,390	10	6,143	6	0.01	0.000	0.0
UA20150148_bag	ARV	Office of the President	UA2015.0148	a2aefaf8-410d-4421-80d4-911c52187a4	None Found	24,831,745	330	26,191	26	0.03	0.000	0.0

Local Administrative Dashboard Figure 10: Gray Repo Admin Console List within List app in Microsoft 365

Admin Console Look and Feel

The default dashboard view, “All Items by Bag#” is a list like in a spreadsheet view sorted by Bag# (or name) as seen in Local Administrative Dashboard Figures 7 and 10 above; however, there are a couple other standard, useful views:

- All Items: Similar to “All Items by Bag#” this presents the data in a tabular view that also sums the total Bits and Files.

Bag	Unit	Collection	AccessionID	FederalID	PII Restrictions	Bits	#Files	Bytes
UA20090117_bag	ARV	Career Connection	UA2009.0117	81cda219-1593-4a01-049a-1aa797b4e699	None Found	40,609,399	168	39,650
UA20150148_bag	ARV	Office of the Provost	UA2015.0148	c8ac9eb-d10d-4e05-8fc6-911c52187a41	None Found	26,821,745	320	26,193
UA20110006_bag	ARV	Faculty Council	UA0211.0006	e556f5-d9c-4176-81fa-0e359eb26871	None Found	2,168,107	7	2,117
UA20110007_bag	ARV	University Senate	UA2011.0007	d34e7a7-b936-40fe-afb9-d45b0d049d4	None Found	6,292,390	10	6,145

Local Administrative Dashboard Figure 11: All Items view in the Gray Repo Admin Console List within Teams

Bag	Unit	Collection	AccessionID	FederalID	PII Restrictions	Bits	#Files	Bytes	Kbs	MBs	GBs	TBs
UA20090117_bag	ARV	Career Connection	UA2009.0117	81cda219-1593-4a01-049a-1aa797b4e699	None Found	40,609,399	168	39,650	39	0.04	0.000	0.0
UA20150148_bag	ARV	Office of the Provost	UA2015.0148	c8ac9eb-d10d-4e05-8fc6-911c52187a41	None Found	26,821,745	320	26,193	26	0.02	0.000	0.0
UA20110006_bag	ARV	Faculty Council	UA0211.0006	e556f5-d9c-4176-81fa-0e359eb26871	None Found	2,168,107	7	2,117	2	0.00	0.000	0.0
UA20110007_bag	ARV	University Senate	UA2011.0007	d34e7a7-b936-40fe-afb9-d45b0d049d4	None Found	6,292,390	10	6,145	6	0.01	0.000	0.0

Local Administrative Dashboard Figure 12: All Items view in the Gray Repo Admin Console List within Microsoft 365 Lists

- **Gallery:** This presents the data in a card/board view with key data initially available on the card.

Bag	Unit	Collection	AccessionID	FederalID	PII Restrictions	#Files	MBs
UA20090117_bag	ARV	Career Connection	UA2009.0117	81cda219-1593-4a01-049a-1aa797b4e699	None Found	168	0.04
UA20150148_bag	ARV	Office of the Provost	UA2015.0148	c8ac9eb-d10d-4e05-8fc6-911c52187a41	None Found	320	0.02
UA20110006_bag	ARV	Faculty Council	UA0211.0006	e556f5-d9c-4176-81fa-0e359eb26871	None Found	7	0.00
UA20110007_bag	ARV	University Senate	UA2011.0007	d34e7a7-b936-40fe-afb9-d45b0d049d4	None Found	10	0.01

Local Administrative Dashboard Figure 13: Gallery view in the Gray Repo Admin Console List within Teams

Bag	Unit	Collection	AccessionID	FederalID	#Files	Bits
UA20090117_bag	ARV	Career Connection	UA.2009.0117	E1c8a219-1593-4a81-8d9a-1aa1...	106	40,609,539
UA20150148_bag	ARV	Office of the Provost	UA.2015.0148	c8ac9eb-d1d-4e05-8fc6-91c5...	320	26,821,745
UA20110006_bag	ARV	Faculty Council	UA.0211.0006	e516fcf5-e9cc-4176-81fa-0e359...	7	2,168,107
UA20110007_bag	ARV	University Senate	UA.2011.0007	d34fe7a7-b936-40fe-ab9-d450...	10	6,292,590

Local Administrative Dashboard Figure 14: Gallery view in the Gray Repo Admin Console List within Microsoft 365 Lists

Entering Data into the Admin Console

While you can enter data in either the List interface in Teams or via the Microsoft 365 List app, we believe it may be slightly easier to do so in Teams.

- The data entry within the Teams Lists interface is more of a landscape layout, allowing you to see all the data in one screen.

Field	Description
Bag	Enter value here
Collection *	Enter value here Name of collection
FederalID *	Enter value here This is the federal ID generated by the ingest process. It is maintained here as well as in the finding aid.
AccessionID	Enter value here The accession ID for the particular "bag". If an accession is split amongst several bags, this may not be a unique ID.
Ingest Documentation *	Enter URL Alternative text Please enter "files" as the default Alternative text for the file type of the ingest document that includes federalID text, pdf text, DROID file, pdf/report and Excel export and possibly a ReadMe file.
Ingested by *	Enter value here Enter first and last name of who ingested the bag into the Gray Repo.
Ingest Date *	Enter a date Date the bag was ingested into the Gray Repo
#Files *	Enter a number The number after the decimal point in the Payload-Count: e.g. if the Payload-Count is 4098.1602-.99, the file count is 40,981,602
Bits *	Enter a number The number before the decimal point in the Payload-Count: e.g. if the Payload-Count is 4098.1602-.99, the bit count is 40,981,602
Notes	Enter value here Enter any necessary or desired notes in this field. Please preference your notes with a date stamp in a 'YYYYMMDD' format. If notes are made on separate dates, enter current notes at the top of field so that they read in reverse chronological order.

Local Administrative Dashboard Figure 15: Data entry screen in Lists within Teams

- Whereas the interface in Lists within Microsoft 365 is more of a portrait orientation that required you to scroll to complete the process.

New item

Bag
Enter value here

Collection *
Enter value here

Name of collection

AccessionID *
Enter value here

The accession ID for the particular "bag". If an accession is split amongst several bags, this may not be a unique ID

PII Restrictions *
 Yes
 None Found

This answers is based upon whether personally identifiable information (PII) was found during the ingest preparation process

Unit *
 ARV
 BPA
 CGA
 GEN
 HIL
 KB
 MUS
 OPPA
 RAR
 TRI

This is the collecting unit that "owns" or is the steward of the

Save **Cancel**

Local Administrative Dashboard Figure 16a: Data entry screen in Lists within Microsoft 365

Ingest Documentation *
Enter a URL

Alternative text

Please enter "Files" as the default Alternative Text. Hyperlink to location of accessioning documentation that includes federalID text, pii text, DROID file, pdf report and Excel export and possibly a ReadMe file.

Ingested by *
Enter value here

Enter first and last name of who ingested the bag into the Gray Repo

Ingest Date *
Enter a date

Date the bag was ingested into the Gray Repo

#files *
Enter a number

The number after the decimal point in the Payload-Oxum; e.g. if the Payload-Oxum = 40981602.169, the file count is 169

Bits *
Enter a number

The number before the decimal point in the Payload-Oxum; e.g. if the Payload-Oxum = 40981602.169, the bit count is 40,981,602

Notes
Enter value here

Enter any necessary or desired notes in this field. Please preference your notes with a date stamp in a "YYYYMMDD" format. If notes are made on separate dates, enter current notes at the top of field so that they read in reverse chronological order.

Save **Cancel**

Local Administrative Dashboard Figure 16b: Data entry screen in Lists within Microsoft 365

- In either interface, there are explanatory notes for each field except for “Bag,” which once again is the bag file name sans the “.zip”.
 - e.g. UA20150148_bag.zip become UA20150148_bag in the Bag field

New item

<input type="text"/> Bag UA20150148_bag	<input type="text"/> Collection * Office of Academic Affairs/Provost	<input type="text"/> AccessionID * UA.2015.0148	<input type="text"/> FederalID * b1ff2af5-8174-48b7-ba28-08b85c0...
Name of collection			
<input checked="" type="radio"/> Unit * ARV	<input type="text"/> Bits * 26891030	<input type="text"/> #files * 319	<input type="text"/> Ingest Documentation * https://buckeyemailosu.sharepoint...
<input type="radio"/> BPA	The number before the decimal point in the Payload-Oxum; e.g. if the Payload-Oxum = 40981602.169, the bit count is 40,981,602		
<input type="radio"/> CGA	The number after the decimal point in the Payload-Oxum; e.g. if the Payload-Oxum = 40981602.169, the file count is 169		
<input type="radio"/> GEN			
<input type="radio"/> HIL			
<input type="radio"/> ...			

Local Administrative Dashboard Figure 177: Data entry comments/instructions in Lists within Teams

- NOTE: Every now and then Lists can have a synching hiccup where the item is not saved and will have to be re-entered. I have encountered this in both versions of Lists. So as a last step, just make sure the entry is in Lists after you have saved it.

What to do with files post-ingest

Now that we have successfully ingested these files into the Gray Repo, as well as established administrative control through the manifests and other files maintained in the _DashboardFiles folders, data in the Local Admin Console and linkages to the finding aid, what do we do with the original files and local bag? Keep in mind the ingest process has us essentially doubling the local storage space by creating the bag. Simply put, we need to practice good records and file management hygiene, and get rid of them in a timely fashion. To that end we need to:

- Rename original files folder: Initially, post-ingest rename the folder with the original files, ingest documentation and the bag in the following manner AccessionID_Ingested_YYYYMMDD (e.g. UA20150148 becomes UA20150148_Ingested_20240112). This provides an additional visual cue that we have completed action with the content contained within the accession. What do we mean by post-ingest? Having successfully received the FederalID text, copied appropriate files to the _DashboardFiles folder, and transcribed the necessary data into the Admin Console.

- Delete:
 - [REDACTED] Bucket: Once you confirmed ingest is complete, have received the Fedoraid text and copied it to the _AccessionID_Ingest_Documentation folder, delete the ingest subfolder in the [REDACTED] Bucket. If something does go wrong with the ingest, you will need to re-copy the bag into the [REDACTED] Bucket anyhow.
 - Original files and bag: Initially, we suggest deleting the original file set and bag no later than two (2) weeks after ingest and receipt of the Fedoraid text file. As we mature with the system we may shrink that window.

Resources

- [Bagger](#)
 - [Bagger GUI User Guide](#) (v [REDACTED] August 2021; State of North Carolina, Department of Natural and Cultural Resources)
- [Bulk Extractor](#) (GitHub site)
 - [User Manual](#) (v1.4 March 23, 2015; Jessica R. Bradley and Simson L. Garfinkel)
- [DROID: file format identification tool](#)
 - [DROID: User Guide](#) (September 2023; National Archives UK)
- [Fedora](#)
- [PRONOM](#)
- [Sustainability of Digital Formats: Planning for Library of Congress Collections](#)