Electronic Records Modules
Electronic Records Committee
Congressional Papers Roundtable
Society of American Archivists

File Transfers Using DataAccessioner

Adriane Hanson
University of Georgia
ahanson@uga.edu

Date Published: December 15, 2016

Module#: ERCM006

Created 2016-07
CPR Electronic Records Committee

Overview and Rationale

When copying files from their original storage environment to the archives, the authenticity and integrity of these files must be preserved. Standard copying procedures can result in errors. Therefore, archivists need to use tools to ensure the copying process was accurate.

DataAccessioner is a free program that compares the MD5 checksum of each file before and after copying to make sure there were no errors, extracts metadata from each file using FITS, and saves the results as a XML file, with most of the metadata following PREMIS.

For more details, see the DataAccessioner Webpage and the DataAccessioner User Guide

Because software mediates and records all interactions, it is possible to have a student assistant or support staff person perform this task. This can free up the time of the digital archivist for more complex work. The digital archivist should monitor their work by checking the DataAccessioner output for errors and may also wish to check a sample of the original disks against the files copied.

Procedural Assumptions

Files are being copied from removable media to a storage location. DataAccessioner can also be run from a thumb drive or external hard drive, so it can be used to copy files from a donor's computer on location. DataAccessioner would not be useful for transferring files over a network.

Removable media has been scanned for viruses so the preservation storage is protected. The media is attached to the computer using write blocking technology to ensure that nothing alters the original files.

This process extracts metadata and automatically generates documentation about the copying process, which is useful for accessioning and preservation monitoring but also slows down the copying process. As a faster alternative, TeraCopy will compare CRC checksums before and after copying to confirm that files are copied correctly but does not extract metadata or generate documentation. Download for free at: http://codesector.com/teracopy

This process is capturing files and maintaining their directory structure, but does not capture any systems information, deleted files, etc. that an original computing environment have. If it is an original computing environment, and factors associated with that environment could be relevant to the archival record, a disk image may be a more appropriate way to copy the donation.

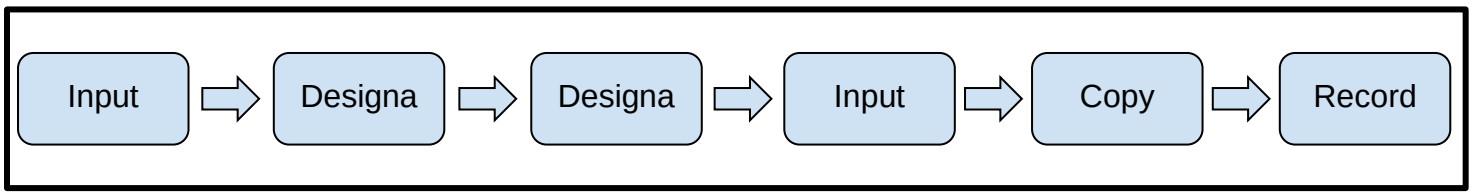
Hardware and Software Requirements

DataAccessioner (v1.0.0-beta) may be downloaded for free at http://dataaccessioner.org/da-version-notes.htm#1.0.0-beta

DataAccessioner requires Java 1.6

To view the XML output as a spreadsheet, also requires the DataAccessioner Metadata Transformer. Download for free at http://dataaccessioner.org/da-mt.htm

Workflow Overview



Workflow

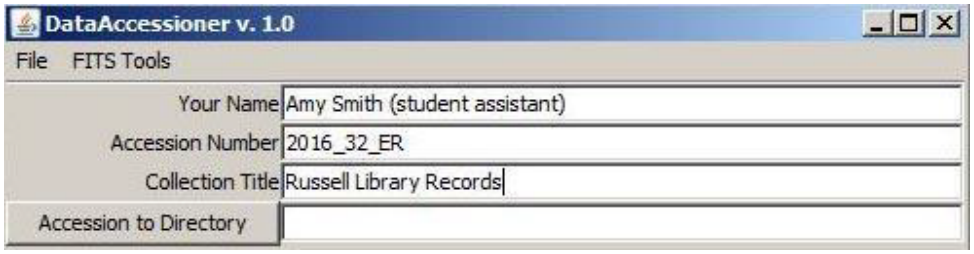
See an example here of the interface completely filled out and ready to copy.

1. **Input Basic Metadata**

   Your Name: include position if not regular staff, i.e. Amy Smith (student assistant)

   Accession #: Format is Year_Sequential Number_ER, i.e. 2016_32_ER
   Note: the "ER" designates that it is electronic records. Papers and other formats received at the same time will have the accession number, i.e. 2016_32.

   Collection Title: Use the official title from the Resource_IDs spreadsheet.



2. Designate Where to Save the Files

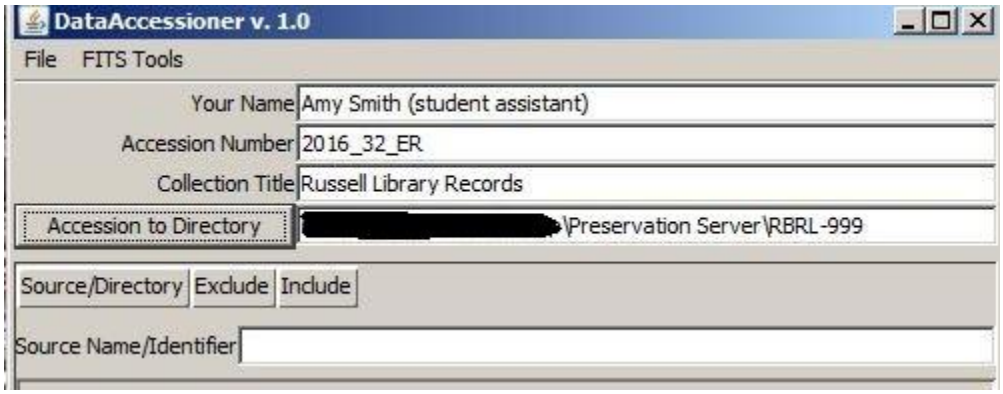
Click on **"Accession to Directory"** and select the folder where the files should be saved. This is the folder named with the collection number located on our preservation server.

During copying, DataAccessioner creates two folders inside the folder designated as the accession directory:

1. Folder named with the accession number, if it doesn't already exist. If it does exist, it will save the files inside the existing accession number folder.

2. Folder named with the disk name (the Source Name/Identifier field) inside the accession number folder.

This means that the accession directory should not be named with the accession number or DataAccessioner will create another folder inside of it also named with the accession number.



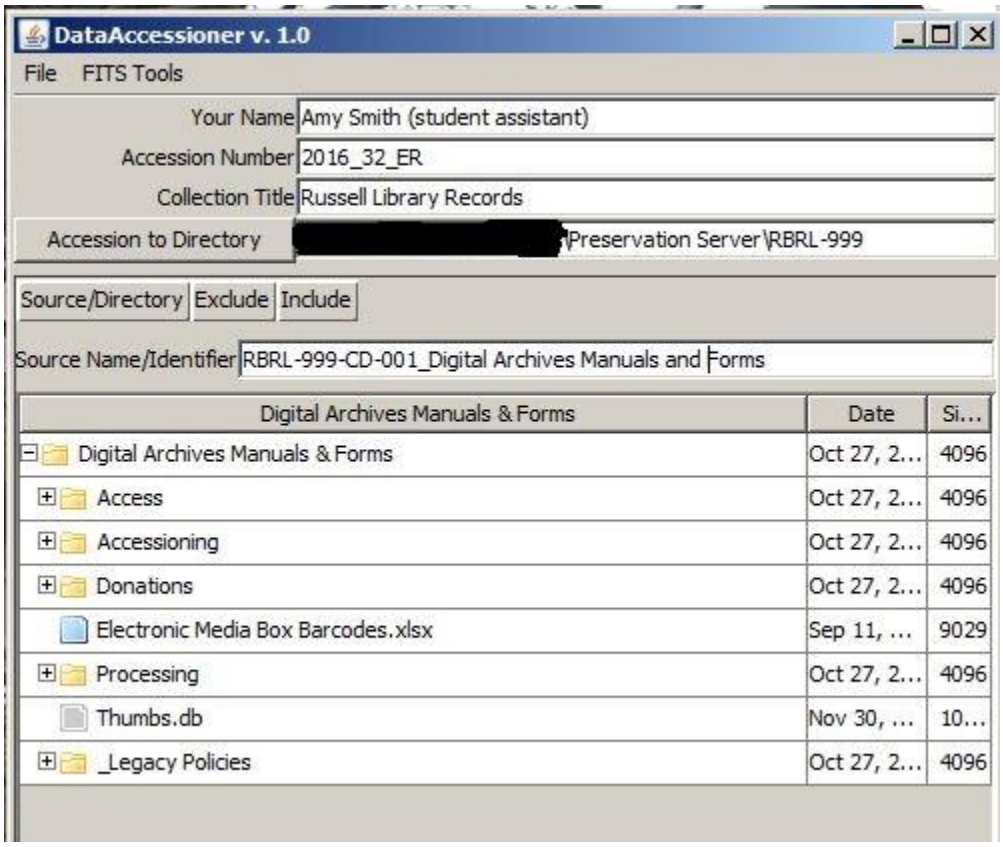
3. Designate Files to Copy

Click on **"Source/Directory**" and select the location of the files. The file directory will be displayed.

Highlight files or folders and click "Exclude" if they should not be copied. Highlight excluded files or folders and click "Include" if they should be copied.

DataAccessioner will populate the field **Source Name/Identifier** with the name that was entered for the disk when it was created. Replace this with the disk name, combining the disk ID number and the disk label. The format is Disk-ID_Disk Name. Replace punctuation and other odd characters (< > : " / \ | ? . ! * ½) with a space, dash or underscore as DataAccessioner will not copy with those characters. Use whatever punctuation best serves to maintain the original meaning of the label.

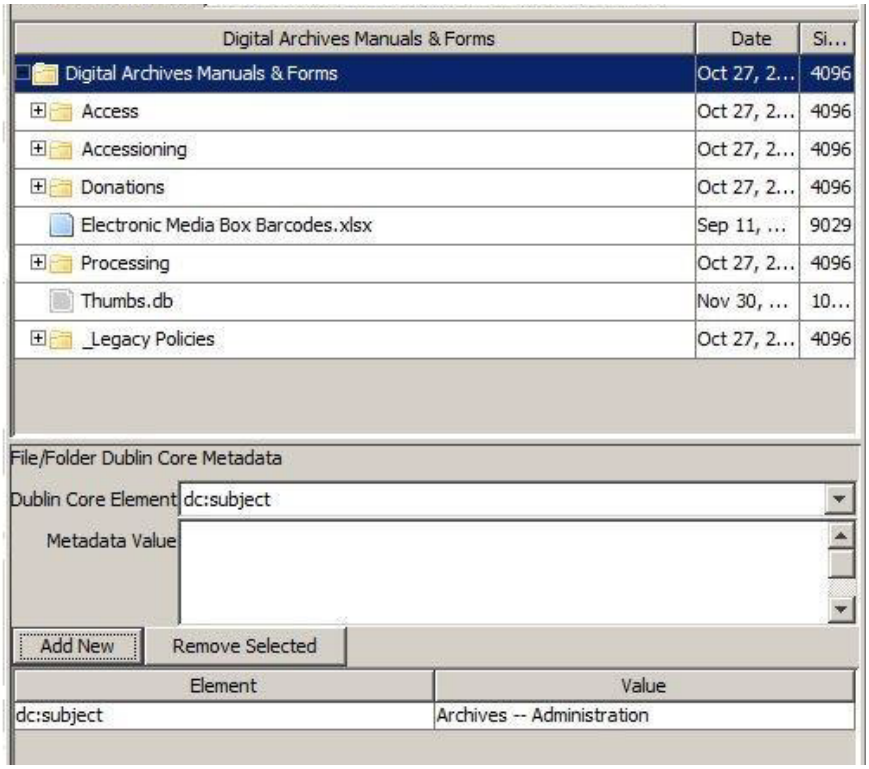
Example: the first CD in collection 350 with the label "Jones Photo: Rome 7/05" is RBRL-350-CD-001_Jones Photo_Rome 7-05



4. Input Dublin Core Metadata (optional)

Highlight the folder that you want to associate the Dublin Core metadata with, usually the top level folder.

Select the Dublin Core Element to add from the dropdown menu and type the value into the Metadata Value box. Then click "Add New". The metadata will become part of the XML output of DataAccessioner.



5. Copy Files

Click "Migrate" to copy the files.

Once migration is complete, if there are no errors, the message "Success!" displays on the bottom left of the interface, under the Migrate button. Otherwise, a window will pop up indicating there are errors and asking if you would like to view them.

Common Errors:

- **Unable to set last modified date: negative time.** Most commonly happens with 3.5" floppy disks. A visual comparison of the files copied and files on the disk generally show that the last modified date is actually unchanged. No action is required other than to document the error.

- **Unable to copy individual files** or **Unable to generate an MD5 sum** for a file. Both errors indicate the file was not copied. Attempt to copy using TeraCopy instead.

- **Unable to detect any files.** Attempt to copy using TeraCopy. If that doesn't work make a disk image using AccessFTK Imager. Also check the disk properties and see if the disk might be empty. If all of those fail, the files are considered unrecoverable and the disk is discarded. Before making this much effort, determine if the files are worth the time to recover based on the disk label and any related papers.

6. Record Results

For each disk, record the results of the copying procedure in the preservation documentation (see example). Standardized text:

> **For no errors:** Copied files from the media to the accession folder on the preservation server with DataAccessioner. No errors were detected.
>
> **For errors:** Copied files from the media to the accession folder on the preservation server with DataAccessioner. The following errors were detected: *list the errors and any actions taken to address those errors*

DataAccessioner generates an XML file (see example) and saves it to collection folder. Rename the file with the disk number and indicate it is from DataAccessioner (i.e. rbrl-350-cd-001_data-accessioner) and save it to the accession folder. This must be done before you copy another disk from that accession or the XML will be overwritten.

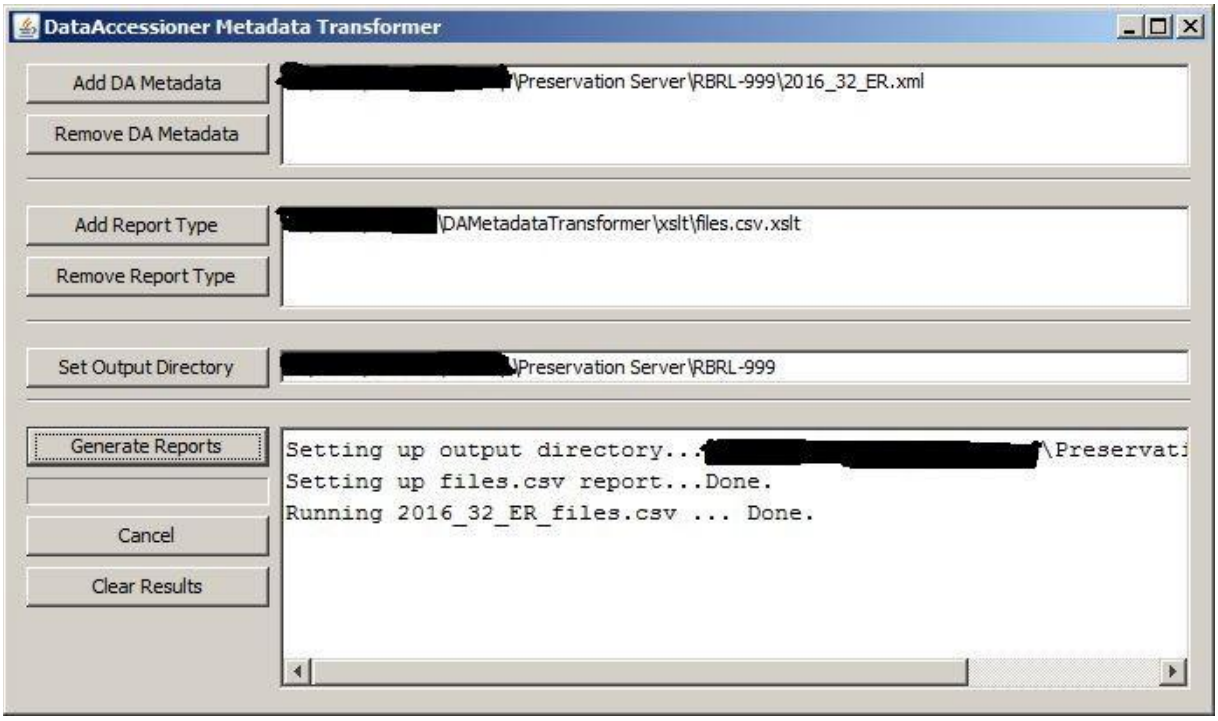
View the XML File as a Spreadsheet

DataAccessioner produces an XML file with metadata about each folder and file, most of it in PREMIS. A subset of this metadata can be easily converted into a spreadsheet using the DataAccessioner Metadata Transformer to make it easier to analyze.

The report contains the Directory Path, File Name, Date Last Modified, Size (in bytes), MD5, and File Format. You could change the XSLT that generates the report to get other fields.

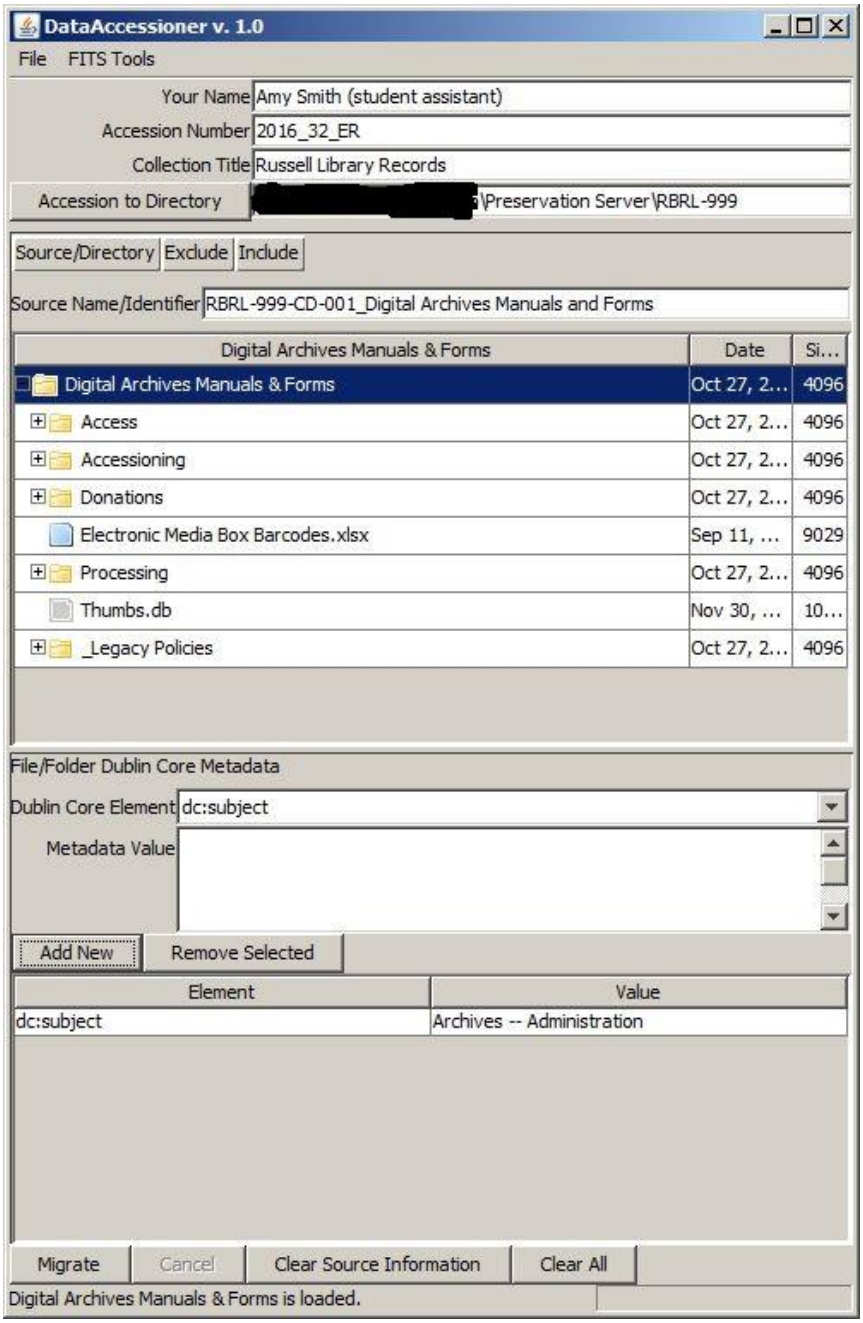
Instructions for generating the report:

1. Click "Add DA Metadata" and select the file you want to convert.
2. Highlight the "files.html.xslt" report and click "Remove Report Type".
3. Click "Select Output Directory" and select where the file should be saved.
4. Click "Generate Reports". This produces a CSV file which can be opened in Excel (see example).



Examples

DataAccessioner Interface - Ready to Copy



Preservation Documentation

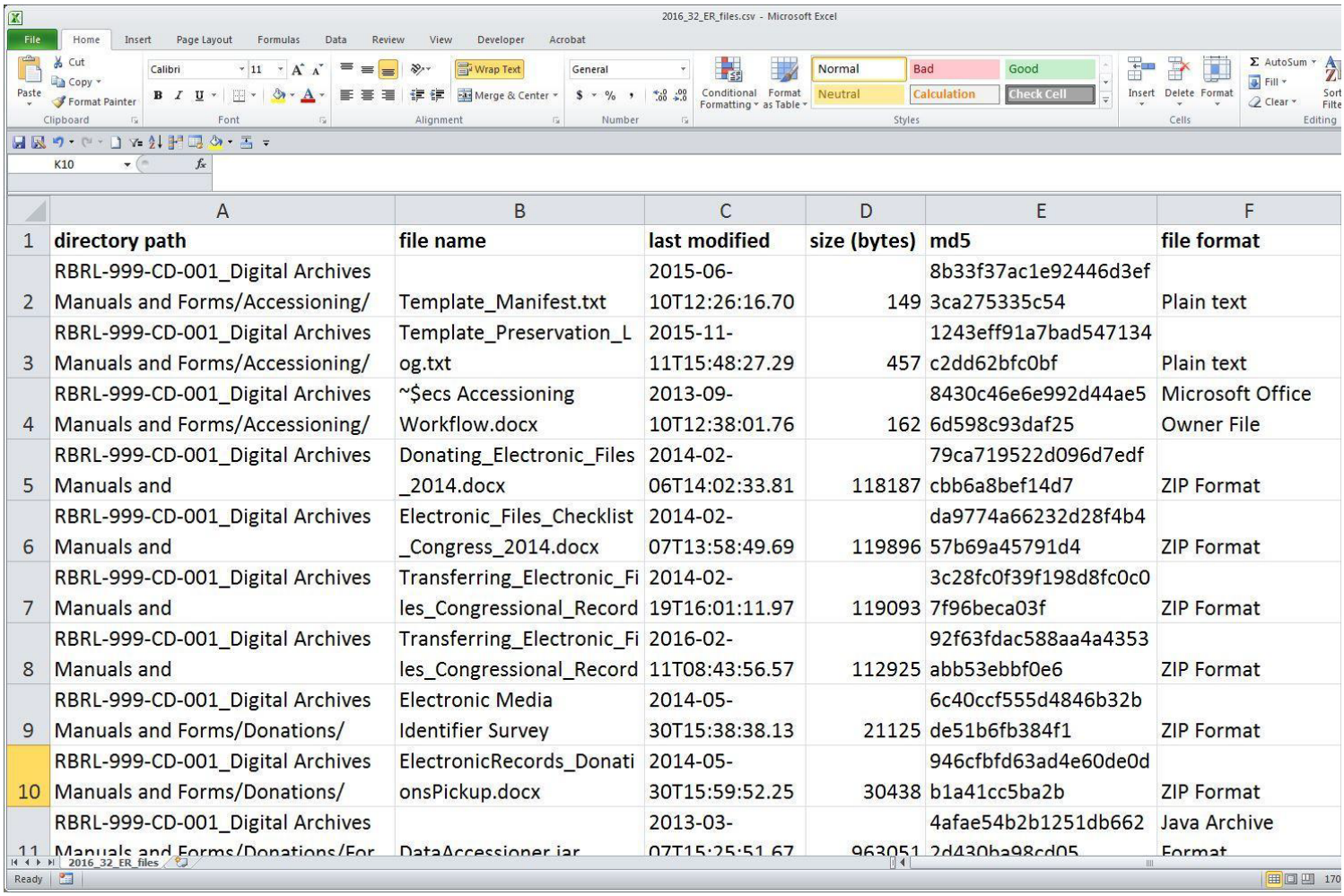
Collection	Accession	Date	Media Identifier	Action	Staff
RBRL-213-ZM	2004-20-ER	2013-11-18	RBRL-213-CD-006	Virus check with Trend Micro OfficeScan 10.5.1997. No security risks were detected.	Rachael Zipperer (student assistant)
RBRL-213-ZM	2004-20-ER	2013-06-03	RBRL-213-CD-006	Copied files from the media to the accession folder on the preservation server with Duke Data Accessioner with the following errors: The file D:\feb-march 2003\022503-24637-AA00.jpg was not found; Migration of D:\feb-march 2003\022503-24637-AA01.jpg interupted. Copied those two files to the accession folder using TeraCopy. No errors were detected.	Adriane Hanson
RBRL-213-ZM	2004-20-ER	2013-11-18	RBRL-213-CD-007	Virus check with Trend Micro OfficeScan 10.5.1997. No security risks were detected.	Rachael Zipperer (student assistant)
RBRL-213-ZM	2004-20-ER	2013-06-03	RBRL-213-CD-007	Copied files from the media to the accession folder on the preservation server with Duke Data Accessioner. No errors were detected.	Adriane Hanson
RBRL-213-ZM	2004-20-ER	2013-11-18	RBRL-213-CD-008	Virus check with Trend Micro OfficeScan 10.5.1997. No security risks were detected.	Rachael Zipperer (student assistant)
RBRL-213-ZM	2004-20-ER	2013-06-03	RBRL-213-CD-008	Copied files from the media to the accession folder on the preservation server with Duke Data Accessioner. No errors were detected.	Adriane Hanson

DataAccessioner XML Output


```xml
<?xml version="1.0" encoding="UTF-8"?>
<collection xmlns="http://dataaccessioner.org/schema/dda-0-3-1" name="Russell Library Records">
  <accession number="2016_32_ER">
    <folder name="RBRL-999-CD-001_Digital Archives Manuals and Forms"
last_modified="2016-10-27T10:12:03.782">
      <dcx:description xmlns:dcx="http://purl.org/dc/xml/">
        <dc:subject xmlns:dc="http://purl.org/dc/elements/1.1/">Archives -- Administration</dc:subject>
      </dcx:description>
      <folder name="Access" last_modified="2016-10-27T09:44:36.604">
        <file name="Digital_Archives_Policy_Access.docx" last_modified="2016-01-29T13:28:01.384"
size="35964" MD5="50b819c4ee1ada335be4397db58b6744">
          <premis:object xmlns:premis="info:lc/xmlns/premis-v2"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:fits="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:uuid="java:java.util.UUID"
xsi:type="premis:file">
            <premis:objectIdentifier>
              <premis:objectIdentifierType>uuid</premis:objectIdentifierType>

<premis:objectIdentifierValue>0d45a9e9-f248-46c4-810f-3edaf4494dce</premis:objectIdentifierValue>
            </premis:objectIdentifier>
            <premis:objectCharacteristics>
              <premis:compositionLevel>0</premis:compositionLevel>
              <premis:fixity>
                <premis:messageDigestAlgorithm>MD5</premis:messageDigestAlgorithm>
                <premis:messageDigest>50b819c4ee1ada335be4397db58b6744</premis:messageDigest>
                <premis:messageDigestOriginator>OIS File Information</premis:messageDigestOriginator>
              </premis:fixity>
              <premis:size>35964</premis:size>
              <premis:format>
                <premis:formatDesignation>
                  <premis:formatName>ZIP Format</premis:formatName>
                  <premis:formatVersion>2.0</premis:formatVersion>
                </premis:formatDesignation>
                <premis:formatNote>Identified by: file utility v5.03</premis:formatNote>
                <premis:formatNote>Identified by: ffident v0.2</premis:formatNote>
              </premis:format>
              <premis:format>
```

DataAccessioner Metadata Transformer - CSV Output



	A	B	C	D	E	F
1	directory path	file name	last modified	size (bytes)	md5	file format
2	RBRL-999-CD-001_Digital Archives Manuals and Forms/Accessioning/	Template_Manifest.txt	2015-06-10T12:26:16.70	149	8b33f37ac1e92446d3ef3ca275335c54	Plain text
3	RBRL-999-CD-001_Digital Archives Manuals and Forms/Accessioning/	Template_Preservation_Log.txt	2015-11-11T15:48:27.29	457	1243eff91a7bad547134c2dd62bfc0bf	Plain text
4	RBRL-999-CD-001_Digital Archives Manuals and Forms/Accessioning/	~$ecs Accessioning Workflow.docx	2013-09-10T12:38:01.76	162	8430c46e6e992d44ae56d598c93daf25	Microsoft Office Owner File
5	RBRL-999-CD-001_Digital Archives Manuals and	Donating_Electronic_Files_2014.docx	2014-02-06T14:02:33.81	118187	79ca719522d096d7edfcbb6a8bef14d7	ZIP Format
6	RBRL-999-CD-001_Digital Archives Manuals and	Electronic_Files_Checklist_Congress_2014.docx	2014-02-07T13:58:49.69	119896	da9774a66232d28f4b457b69a45791d4	ZIP Format
7	RBRL-999-CD-001_Digital Archives Manuals and	Transferring_Electronic_Files_Congressional_Record	2014-02-19T16:01:11.97	119093	3c28fc0f39f198d8fc0c07f96beca03f	ZIP Format
8	RBRL-999-CD-001_Digital Archives Manuals and	Transferring_Electronic_Files_Congressional_Record	2016-02-11T08:43:56.57	112925	92f63fdac588aa4a4353abb53ebbf0e6	ZIP Format
9	RBRL-999-CD-001_Digital Archives Manuals and Forms/Donations/	Electronic Media Identifier Survey	2014-05-30T15:38:38.13	21125	6c40ccf555d4846b32bde51b6fb384f1	ZIP Format
10	RBRL-999-CD-001_Digital Archives Manuals and Forms/Donations/	ElectronicRecords_DonationsPickup.docx	2014-05-30T15:59:52.25	30438	946cfbfd63ad4e60de0db1a41cc5ba2b	ZIP Format
11	RBRL-999-CD-001_Digital Archives Manuals and Forms/Donations/For	DataAccessioner.jar	2013-03-07T15:25:51.67	963051	4afae54b2b1251db6622d430ba98cd05	Java Archive Format