

The guides on the following pages were used during the hands-on session, Demystifying Digital Records Processing: Step-by-step, Byte-by-Byte, at the 2015 Library Technology Conference.

Demystifying Digital Records Processing

Step-by-Step, Byte-by-Byte

Presented at: Library Technology Conference, St. Paul, MN. March 19, 2015



LIBRARIES

Carol Kussmann, Digital Preservation Analyst
Lara Friedman-Shedlov, Description & Access Archivist

Image: James Woodward - jameswoodward.wordpress.com/2010/05/07/mist/

Thank You

U of M Libraries Electronic Records Task Force Web Page
z.umn.edu/libertf

U of M Libraries Digital Preservation Web Page
www.lib.umn.edu/dp/guides
www.lib.umn.edu/dp/glossary

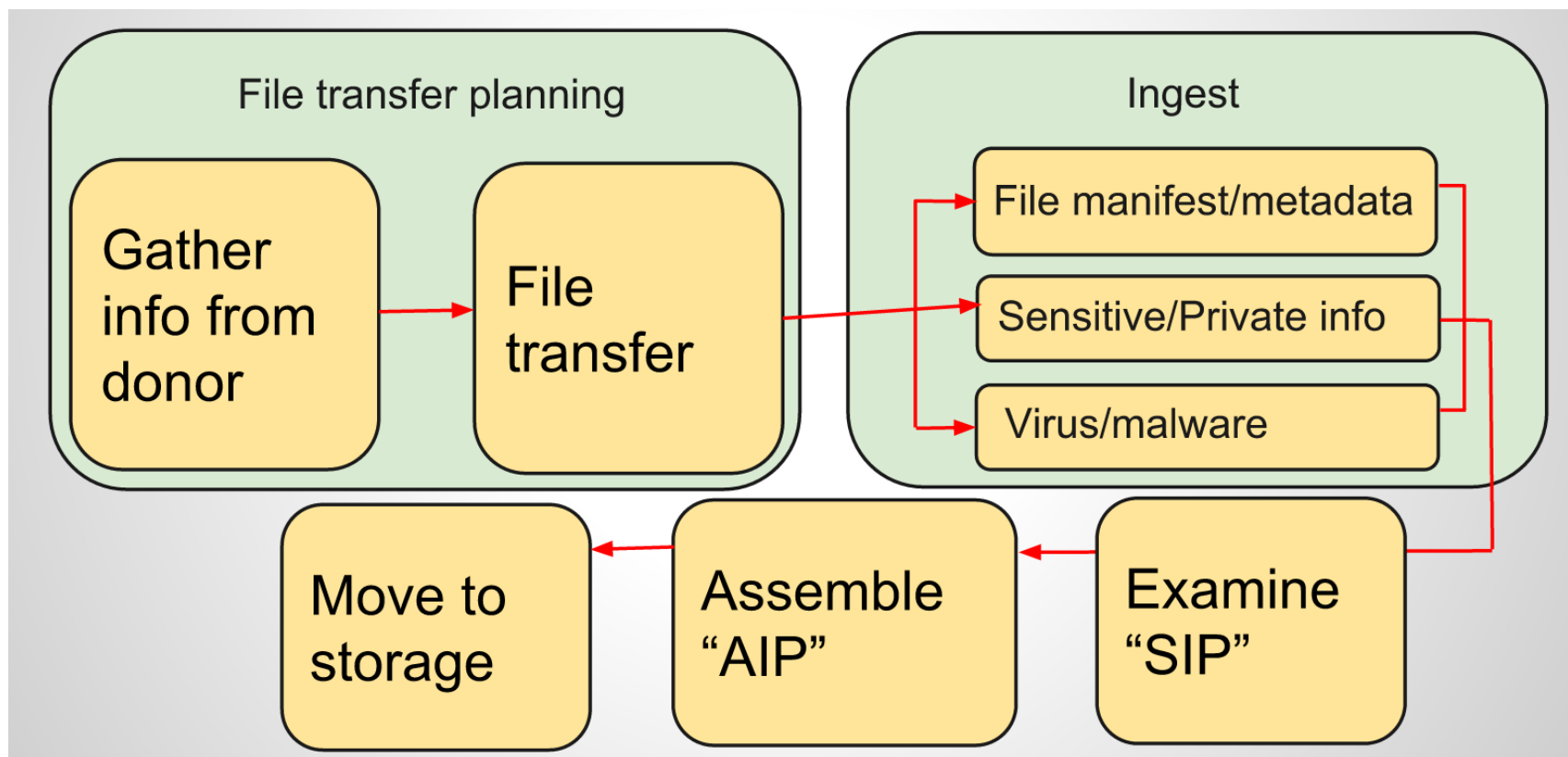
Lara Friedman-Shedlov | ldfs@umn.edu | @ldfs

Carol Kussmann | kussmann@umn.edu



Basic Electronic Records Processing Workflow

UMN Libraries - Electronic Records Task Force



Quick Reference Guide for Tools to Manage and Protect Your Digital Content

The following chart lists and describes many tools that may be helpful when working with digital files. More information on these tools and/or step-by-step guides on how to use them can be found on the University of Minnesota Libraries Digital Preservation Page (www.lib.umn.edu/dp/guides).

Updated March 2015.

Tool Name	Used For	Description	Windows Download/Alternative Program	Mac OS Download/Alternative Program
Bulk Rename Utility	Renaming Files	A program for renaming files in bulk. File names can be appended, prepended, or changed - specific settings allow users to choose how this happens.	z.umn.edu/okv	No Mac OS download Alternative Program: Name Changer: z.umn.edu/okw
Data Accessioner	Data Transfer	Transfers files from one location to another. Uses MD5 checksums to verify the process was successful. In addition to copying files, an XML file about the transfer is generated that documents associated (entered) descriptive metadata and collected technical metadata about the files.	dataaccessioner.org (java based)	dataaccessioner.org (java based)
Disk Inventory X	File Visualization	Helps visualize the contents of selected drives and learn more about the properties of the content, and allows for deletion of content.	No Windows download Alternate program: windirstat.info	derlien.com Alternate Program: GrandPerspective: grandperspectiv.sourceforge.net
DROID	Checksum (MD5) File Format Identification	Generates MD5 checksums. Performs batch identification of file formats, and links to central registry for technical information (PRONOM). Generates reports (PDF), or saves results as .CSV. Helps to understand your digital assets, assess risk, and plan for action (migration to archival file formats, for example).	z.umn.edu/k68	z.umn.edu/k68
Duplicate File Finder	Duplication Detection	Find and delete duplicate files. Can compare content byte by byte or match on file name.	ashisoft.com	No Mac OS download Alternative Program: Araxis Find Duplicate Files: z.umn.edu/k6c A search in the App Store for "file duplicates" returns both free and paid options.

Quick Reference Guide for Tools to Manage and Protect Your Digital Content

Tool Name	Used For	Description	Windows Download	Mac OS Download/Alternative Program
Exact File	Checksum (<i>MD5, SHA1, SHA256, & more</i>)	Helps maintain file integrity: Use to make sure files haven't been changed or damaged over time, or in the process of moving files from one location to another.	exactfile.com	DROID captures MD5 checksums. A search for MD5, Sha-1, or Hash in the App Store brings up many free and paid options.
Hash My Files	Checksum (<i>MD5, SHA1, CRC32, SHA-256, & more</i>) Duplication Detection	Helps maintain file integrity: Use to make sure files haven't been changed or damaged over time, or in the process of moving files from one location to another. Saves checksum values as CSV/Tab-Delimited.	z.umn.edu/k69	DROID captures MD5 checksums. A search for MD5, Sha-1, or Hash in the App Store brings up many free and paid options. A search for "file duplicates" also returns many results.
Karen's Directory Printer	Checksum (MD5) File Format Identification	Helps maintain file integrity by generating MD5 checksums. Prints names of every file on a drive, file size, date/time of last modification, and file attributes. Saves information as .txt file.	z.umn.edu/k6a	DROID captures file format information, selected file/folder metadata, and can capture MD5 checksum values.
Metadata Quality Control (MDQC)	Metadata Quality Control	Compare <i>expected</i> technical (and some non-technical) metadata to <i>actual</i> technical metadata. (E.g. run the tool on a group of files returned to you from a vendor).	avpreserve.com/tools/mdqc	avpreserve.com/tools/mdqc
Name Changer	Rename Files	A program for renaming files in bulk using batch processing. File names can be appended, prepended, or changed - specific settings allow users to choose how this happens.	Alternate Programs: Bulk Rename Utility and ReNamer	[Update: March 2015, link to download not working. See guide for original link.]
NARA File Analyzer	Analyze folder/drive contents. Metadata collection.	Analyze the contents of a folder/drive and export results as .txt file. (E.g. count number of files by type, create directory listing, document checksums)	z.umn.edu/tbk	
Remove Empty Directories	File Clean Up	Identifies and is able to delete empty directories within a chosen location.	z.umn.edu/oky	
ReNamer	Rename Files	A program for renaming files in bulk using batch processing. File names can be appended, prepended, or changed - specific settings allow users to choose how this happens.	z.umn.edu/uft	No Mac OS download Alternative Program: Name Changer: z.umn.edu/okw

Quick Reference Guide for Tools to Manage and Protect Your Digital Content

TeraCopy	Move / Copy Files	Moves or copies files from one location to another without changing the 'created date' of the file. File transfers can also be verified.	z.umn.edu/ss2	
Windows Directory Statistics (WinDirStat)	File Visualization	Helps visualize the contents of selected drives and learn more about the properties of the content, and allows for deletion of content.	windirstat.info	Alternative Programs: Disk Inventory X: derlien.com GrandPerspective: grandperspectiv.sourceforge.net

Data Accessioner Resource Guide



DataAccessioner.jar

Tool Name: Data Accessioner

Functionality/Use: Data Accessioner is a tool that can be used to move content from one location to another. During this process MD5 checksums are created and recorded in the XML output file. Other information captured in the XML file includes file format identification, last modified date, and file size. (Platform independent)

Note: This was written based on v1.0.0 beta. User experience may vary as new versions and updates are made.

To Install:

- Download Data Accessioner (v1.0.0 beta) from the DataAccessioner website: <http://dataaccessioner.org/> (DataAccessioner_v1_0-beta.zip)
- Unzip the folder and move to a location of your choice.

To Use:

- Open the DataAccessioner_v1_0-beta folder and double click on the DataAccessioner.jar file. This opens up the main window.

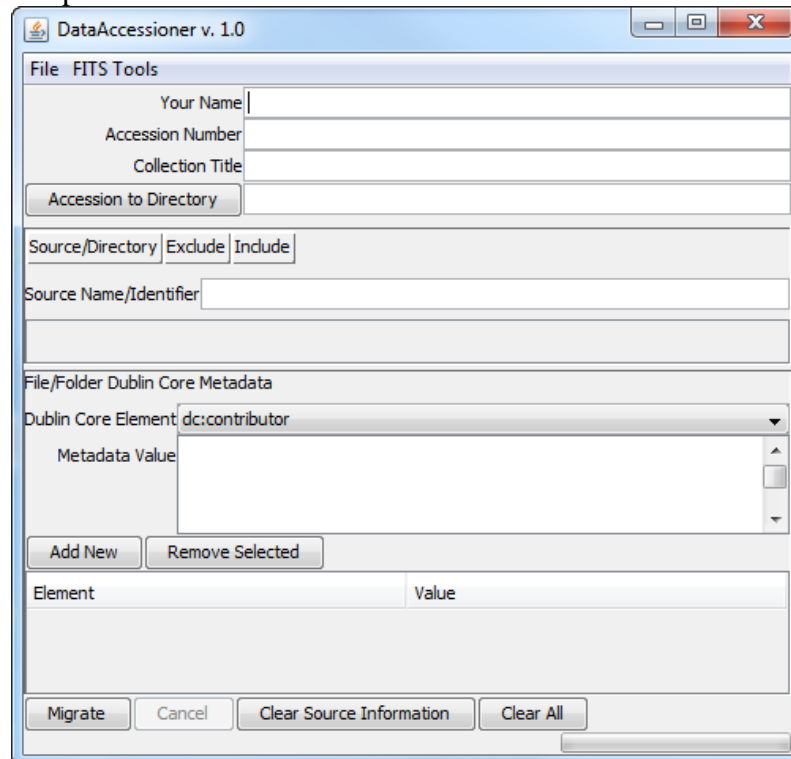


Figure 1: Main Window for Data Accessioner

- Click the FITS Tools dropdown menu and verify that all (or your desired) options are selected.

- Enter in the information about the file transfer/accession.
 - o Your Name: Often the name of the person completing the transfer
 - o Accession Number: Often the accession number for the collection. (This becomes a new folder in the selected location.)
 - o Collection Title: Often the name of the collection.
 - o *Note: The information you provide in these three fields may vary based on your organizational needs – this may be especially true for accessions where more than one transfer is necessary. To determine this, review Figures 2, 3, and 4 and do some experimenting of your own.*

- Select the location in which you want the files to be copied to by clicking the “Accession to Directory” button.

- Identify the location of the files you want to transfer by clicking “Source/Directory”. This information will populate the “Source Name/Identifier” field.

- Change the name of the “Source Name/Identifier” field if necessary.
 - o Special characters, such as : or / are not accepted; which is often included in many file paths names. If you click migrate before changing this name, you will receive an error message.
 - o *Note: Data Accessioner uses this name to create a folder with the same title in the selected location.*

- Review the list of files that appear after selecting the “Source/Directory”.
 - o All files will be transferred by default
 - o To select files that will NOT be transferred, highlight the file, and click the “exclude” button. This will put a red stop sign on the file indicating that it will not be transferred. If you change your mind, highlight the same file and click “include” and the file will once again be included in the files to transfer.

- Add additional metadata to the overall accession or to individual files if desired.
 - o Highlight the file or folder level to add metadata to.
 - o Choose the dc: metadata element from the Dublin Core Element dropdown list
 - o Enter in the Metadata Value in the open box
 - o Click the “Add New” button.
 - o Continue this process until you are done adding metadata.

- If you would like to remove metadata, highlight the metadata element you no longer want to associate with the accession and click the “Remove Selected” button.
- When all of the information is complete and you are ready to move the files, click the “Migrate” button. The files are then copied from the original location to the location you selected in the “Accession to Directory” field.

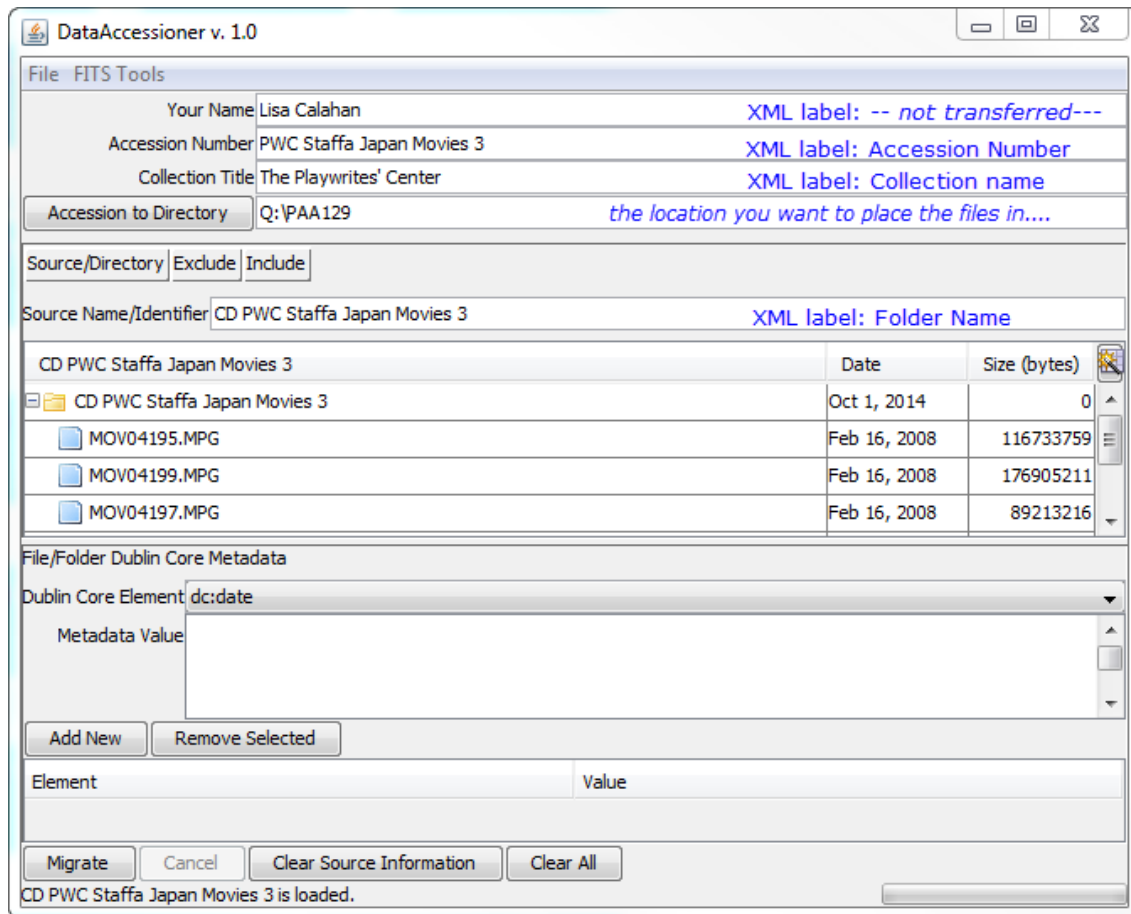


Figure 2: Screenshot of completed data entry with indicators of where the information will be in the resulting XML file.

Results:

- File Transfer
 - Data Accessioner uses the information you provided to copy the files and place them within a nested folder structure. In the example below:
 - *Q* is the drive location and *PAA129* folder was created prior to opening Data Accessioner.

- The *PWC Staffa Japan Movies 3* folder was entered as the “Accession Number” which Data Accessioner used to create a new folder.
- The *CD PWC Staf..an Movies 3* folder was entered as the “Source Name/Identifier” which Data Accessioner used to create a new folder.

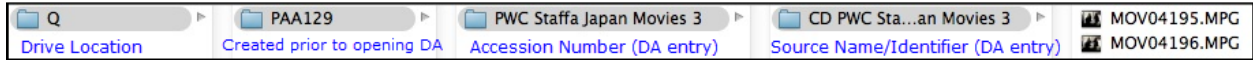


Figure 3: Folder structure created for files transferred with Data Accessioner

- XML File
 - o An XML file is created documenting the transfer and recording specified descriptive metadata and generated technical metadata about the files that were transferred.
 - o The following screenshot shows how the descriptive and some technical metadata appears in the file.

```
<?xml version="1.0" encoding="UTF-8"?>
- <collection name="The Playwrites Center" xmlns="http://dataaccessioner.org/schema/dda-0-3-1">
  - <accession number="PWC Staffa Japan Movies 3">
    - <folder name="CD PWC Staffa Japan Movies 3" last_modified="2008-02-17T18:41:51.000">
      - <file name="MOV04195.MPG" last_modified="2008-02-16T10:51:58.000" MD5="f1af0dfd44
```

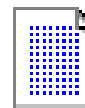
Figure 4: XML file with the descriptive and technical metadata.

Things to Consider

- Data Accessioner can't transfer .cda (CD audio) files. When looking at the transfers, the file name is generated in the new location, but the file itself is not transferred. No error messages are generated.
- Data Accessioner functionality could change with new versions (so functionality may change).
- Data Accessioner is dependent on the FITS Tool, which in turn uses multiple tools – all of which may change over time. (Who is in charge of making sure things are up to date?) During initial testing, it was noticed that the DROID signature file included in Data Accessioner, which was pulled from FITS was older than what DROID as a stand-alone tool was using. When run against each other and compared different results were obtained. Dependencies on other tools may or may not be an issue, but is something to be aware of.

More information (older version): <http://e-records.chrisprom.com/dataaccessioner/>

Hash My Files Resource Guide



Tool Name: Hash My Files

Functionality/Use: Create checksum value/s for monitoring files over time. The program can calculate multiple checksums at the same time. The program can also identify duplicate files. (Windows only)

To Install:

- Download at NirSoft: http://www.nirsoft.net/utils/hash_my_files.html (the download link can be found 2/3 of the way down the page in light purple)
- Unzip 'hashmyfile.zip'

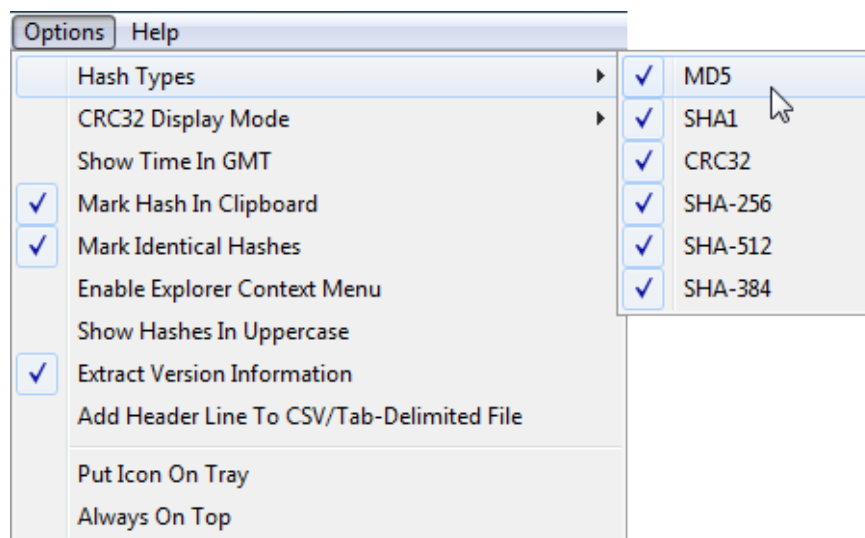
To Use:

- Open the hashmyfile folder and double click HashMyFile.exe

Selecting Hash Types and Other Options

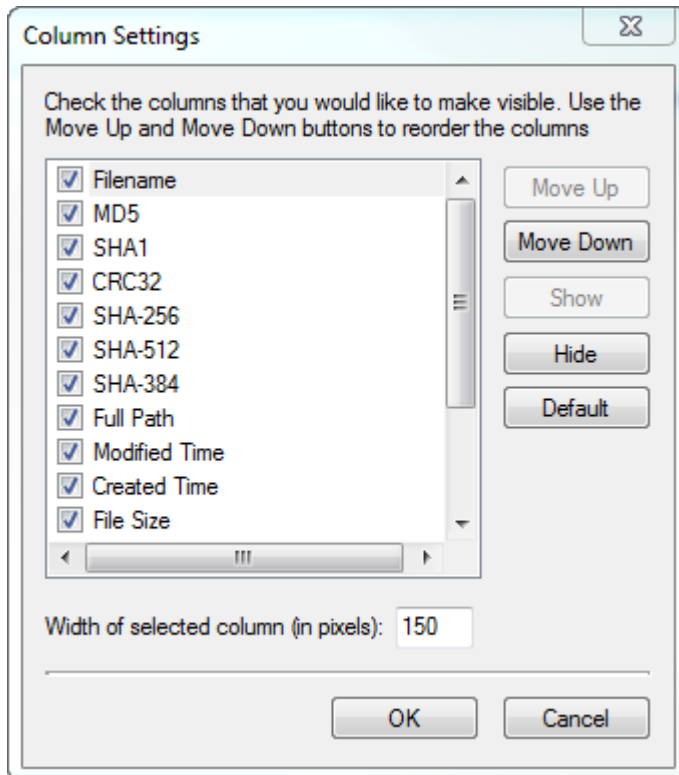
Before you run the program, you may want to take time and choose your settings.

- Under Options, go to Hash Types, and select only the checksum algorithms you want to use. You can only uncheck one box at a time; you will have to navigate to the menu for each one you want to select or remove.
- To take advantage of identifying duplicate files, make sure the "Mark Identical Hashes" is checked.



Choosing Columns to View and for Reports

- Under View, go to Choose Columns, then select the items you want to include on the screen as well as in a CSV report/file. You can also order the columns here by highlighting the item you want to move and click the Move Up or Move Down buttons appropriately.



- The resulting header bar might look like this:

Filename	MD5	Extension	File Size	Identical	File Attributes
----------	-----	-----------	-----------	-----------	-----------------

- After the settings are the way you want them to be you can select files on which to calculate hash values for.

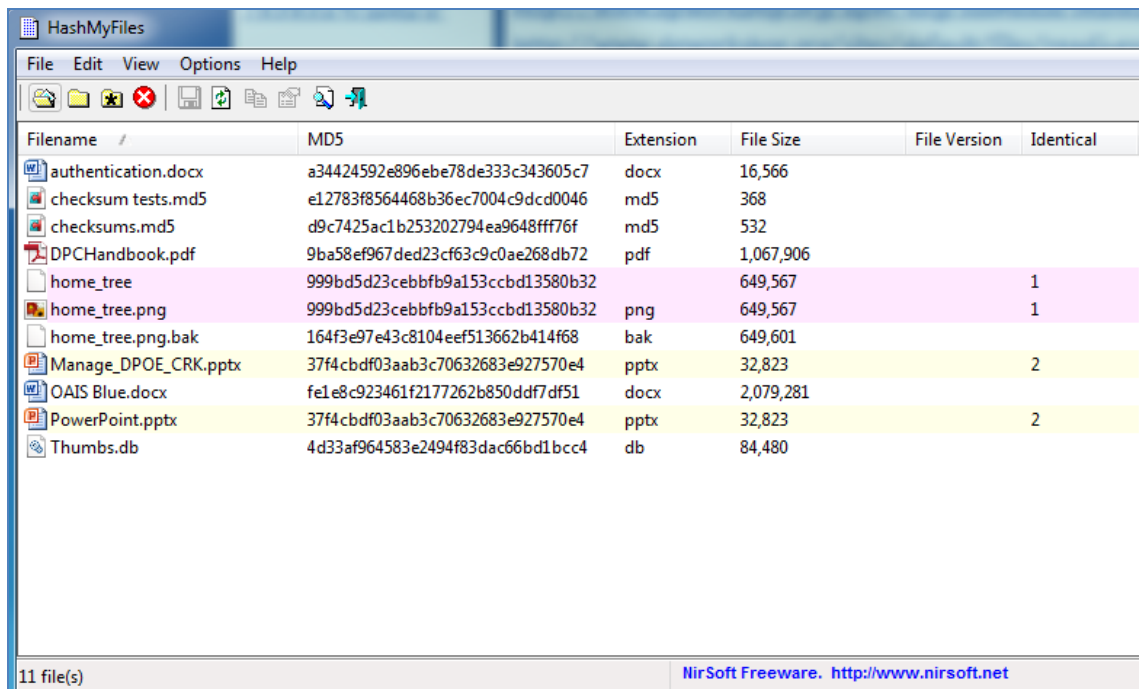
Calculate Hash Values

[There are many ways to do this, only one way is shown below.]

- Go to File, Add Folder
- Browse to the folder you want to add
- Select Add Files in Subfolders to capture any folder hierarchy
- Click OK

The checksums will be calculated and information collected displayed. One added benefit of this program is it visually demonstrates if duplicates are found by color coding and labeling the duplicate files.

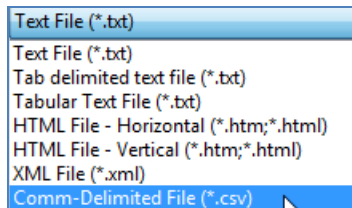
All headers can be used to sort the information.



Save a Report

- Sort the data and move the column order around if you choose.
- [If you want to include headers in a csv file or tab delimited text file, you will need to click the box indicating your preference under Options.]
- Highlight all of the records you want to include in the report (“Ctrl A” on your keyboard to select all items.)
- To save click the Disk Icon or use the File menu and select Save Selected Items.
- In the Save Window

- o Select the location you want to save the file in
- o Name the report
- o Choose a format



- o Click Save

A CSV report looks something like this:

	A	B	C	D	E	F	G
1	Filename	MD5	Extension	File Size	File Version	Identical	
2	authentication.docx	a34424592e896ebe78de333c343605c7	docx	16,566			
3	checksum tests.md5	e12783f8564468b36ec7004c9dcd0046	md5	368			
4	checksums.md5	d9c7425ac1b253202794ea9648fff76f	md5	532			
5	DPCHandbook.pdf	9ba58ef967ded23cf63c9c0ae268db72	pdf	1,067,906			
6	home_tree	999bd5d23cebbfb9a153ccbd13580b32		649,567			1
7	home_tree.png	999bd5d23cebbfb9a153ccbd13580b32	png	649,567			1
8	home_tree.png.bak	164f3e97e43c8104eef513662b414f68	bak	649,601			
9	Manage_DPOE_CRK.pptx	37f4cbdf03aab3c70632683e927570e4	pptx	32,823			2
10	OAIS Blue.docx	fe1e8c923461f2177262b850ddf7df51	docx	2,079,281			
11	PowerPoint.pptx	37f4cbdf03aab3c70632683e927570e4	pptx	32,823			2
12	Thumbs.db	4d33af964583e2494f83dac66bd1bcc4	db	84,480			
13							
14							

Another guide with additional information can be found here:

http://www.mnhs.org/preserve/records/legislativerecords/carol/docs_pdfs/HashMyFilesEvaluation.pdf

Duplicate File Finder Resource Guide



Tool Name: Duplicate File Finder

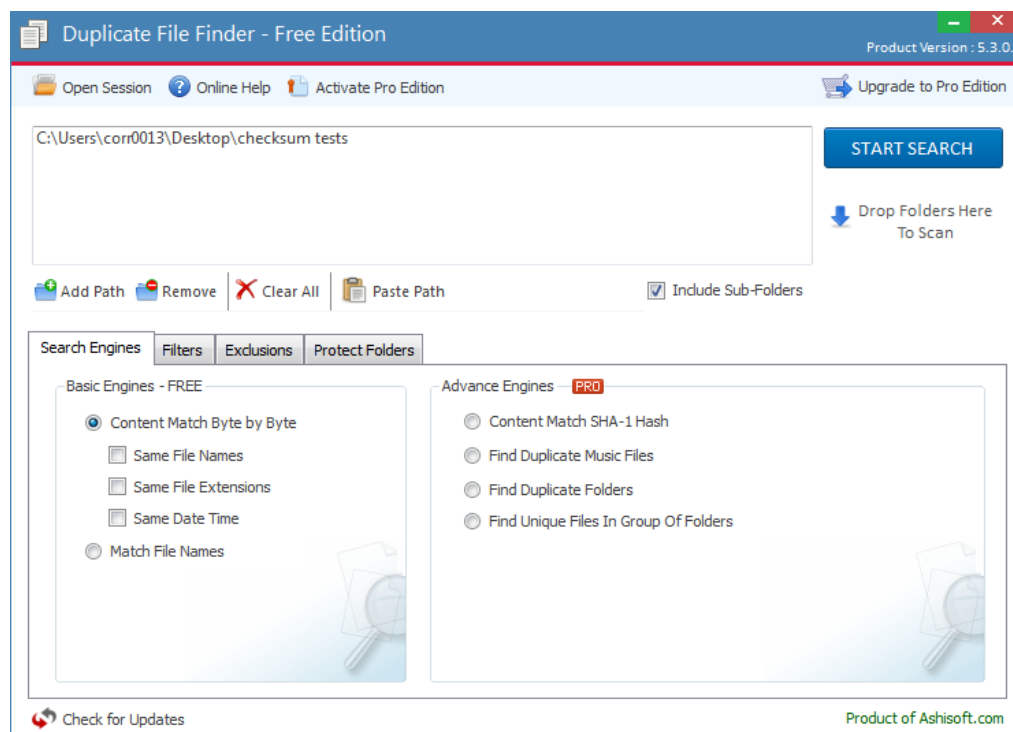
Functionality/Use: Find duplicate files in a selected group of files; you can also choose to immediately delete the files as well. [The free version allows you to find and delete files, but additional features are available in the Pro (paid) Version.] (Windows only)

To Install:

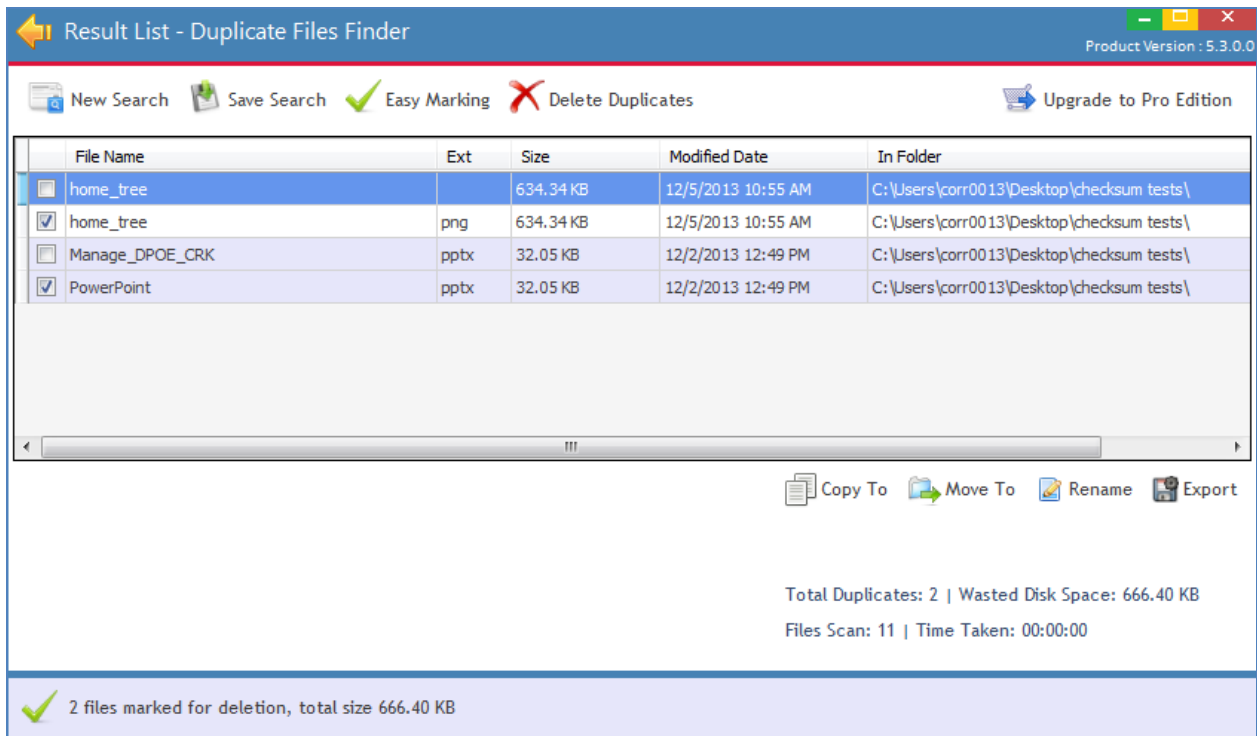
- Download from Ashisoft at: <http://www.ashisoft.com/> (dfsetup.exe)
- Double click file (dfsetup.exe) and select Run
- Type in Admin User Name and Password
- Walk through Setup Wizard

To Use:

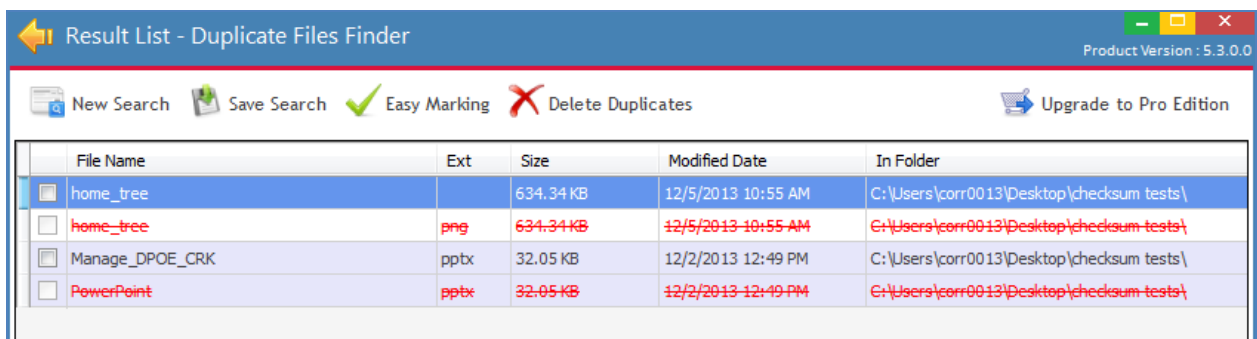
- Double click on program Icon
- Choose the area/s that you want to search for duplicate files (Add Path)
- Choose the method in which to find duplicate files (“Content Match Byte by Byte” is most thorough)
- Click Start Search



- Results will be returned



- Verify files to be deleted (if you want to delete a different file, just check/uncheck the boxes next to the file name). Files with the checkmark will be deleted.
- Delete Duplicates (send to recycle bin or delete immediately)



Additional Resources for More Information

- Online Tutorial from Company: <http://www.ashisoft.com/find-duplicates-step-by-step.htm>
- Guide: http://www.mnhs.org/preserve/records/docs_pdfs/DuplicateFileFinder.pdf