FILE DISCOVERY COMMAND LINE SCRIPT

DESCRIPTION:

Analyzes directories of objects and records, and prepares a csv file for preservation/ingest. This prepares a file so the curators can work with the materials and select those that need to be preserved, and then add collection and metadata information. Complement to the csv ingest tool.

When the program is run, the resulting file will be named: fileList_*[datetime]*.xls.
For example: fileList_201412051315.xls
The default setting is that the file is written to the folder that was scanned

The file headers are:
FILENAME, ITEM ID, FILEPATH, BYTESIZE, SIZE, COLLECTION, IE_LEVEL, DATE_CREATED, DATE_MODIFIED, TITLE, CREATOR, DESCRIPTION, RIGHTS_POLICY

On Windows, with Excel 2007 or greater, the maximum number of rows is 1,048,576; previous versions had a maximum of 65536 rows. This utility works with the higher row limit.
If there are more items than you want in the spreadsheet, this script can also be run on a path you designate. You can run the script on a sub-directory that will fit the number of items you want to work with.

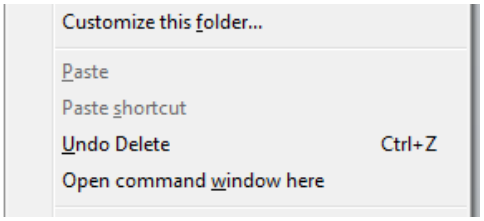
RUNNING THE SCRIPT:

To run the script, you can either

a. copy the FileDiscovery.jar file from the …\DigitalPreservation\Tools folder to your computer, or
b. you can run the script from that folder location.

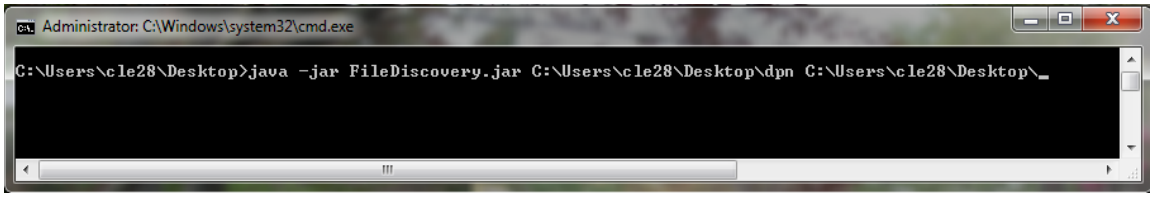
This script is intended to be run on a Windows workstation. It must be run from the command prompt. To open a command prompt on a Windows computer, you can either:

1. Enter *cmd.exe* at the Start Menu search, or
2. SHIFT+Rightclick on a folder and selecting “Open command window here”.



The folder that you click on should be the same folder where the FileDiscovery.jar file resides.

From the command prompt, enter:
java -jar FileDiscovery.jar [path name to check] [output path name for saving the report]



NOTE: If you run the command without specifying the output path, the report will be saved in the folder being checked. This may not be desirable for archival folder that should not be changed.

If you are running the script a number of times against different folders, it may be easiest to run the FileDiscovery.jar file from the same location. In that case, the command to run the script needs to have the correct path for the jar file, so the command could look like this:
java -jar **C:\Tools**FileDiscovery.jar C:\Users\userid\Desktop\dpn C:\Users\userid\Desktop\

Examples:
java -jar FileDiscovery.jar **g:** [for an entire directory]
java -jar FileDiscovery.jar C:\Users\userid\Desktop\dpn C:\Users\userid\Desktop\
java -jar **C:**FileDiscovery.jar C:\Users\userid\Desktop\dpn C:\Users\userid\Desktop\
Add the path name where the FileDiscovery.jar file is located if you want to have it in a permanent folder that you run it from.
You can run the program from the M drive by using the correct path:
java -jar **...\DigitalPreservation\Tools**FileDiscovery.jar

Run file:
java -jar …\DigitalPreservation\Tools\FileDiscovery.jar C:\Users\userid\Desktop\OrsonScottCardArchive C:\Users\userid\Desktop\

java -jar …\DigitalPreservation\Tools\FileDiscovery.jar R:\cdm\objects C:\Users\userid\Desktop\

java -jar ...\DigitalPreservation\Tools\FileDiscovery.jar C:\Users\userid\Desktop\testia\Sources\CAMPTifFiles C:\Users\userid\Desktop\

Tips:
1. If you are copying the command from a Word document, make sure that the character is a - (hyphen) and not a – (dash) which will cause an error.
2. If the path refers to a folder with a space in the name, put the entire path in quotes: "C:\Users\userid\Desktop\Digital Preservation"

PREPARING THE SPREADSHEET FOR PROCESSING FOR ROSETTA:

Currently the harvest utility does not work with multiple folders. Until that is ready, this script can only be used for information purposes.

The final spreadsheet must be reformatted before it can be submitted to Rosetta for harvesting. The following steps need to be completed:

Edit the spreadsheet:

- Add Row 1 with source file names
- Add any additional metadata columns you wish to include
- Update COLLECTION column with the name of the collection in Rosetta
- FILEPATH and IE_LEVEL: not implemented yet; remove columns
- Update TITLE, CREATOR, DESCRIPTION, RIGHTS_POLICY columns
- Remove BYTESIZE and SIZE columns
- Check the file names: The file names must **not** contain spaces or hyphens. If so, you must change both the file name and the file name entry on the spreadsheet. Otherwise, it will result in an error when ingesting into Rosetta.

If you have questions, please ask for assistance.