File Inventory with DROID

Updated January 2018

Tool Homepage: http://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

Introduction

The Digital Record Object Identification (DROID) File Profile Tool was developed by The National Archives (of the UK) to perform automated identification of file formats. It uses *digital signatures* to identify the file format and version beyond what can be identified by the file extension alone. It uses information from the PRONOM database (also managed by the National Archives) found here: http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

DROID is free and open source under the New BSD License.

There is an extensive user guide for the GUI version as well as the tools available via the command line. http://www.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf

Below is an example workflow of file inventorying at the American Geographical Society Library at UWM Libraries.

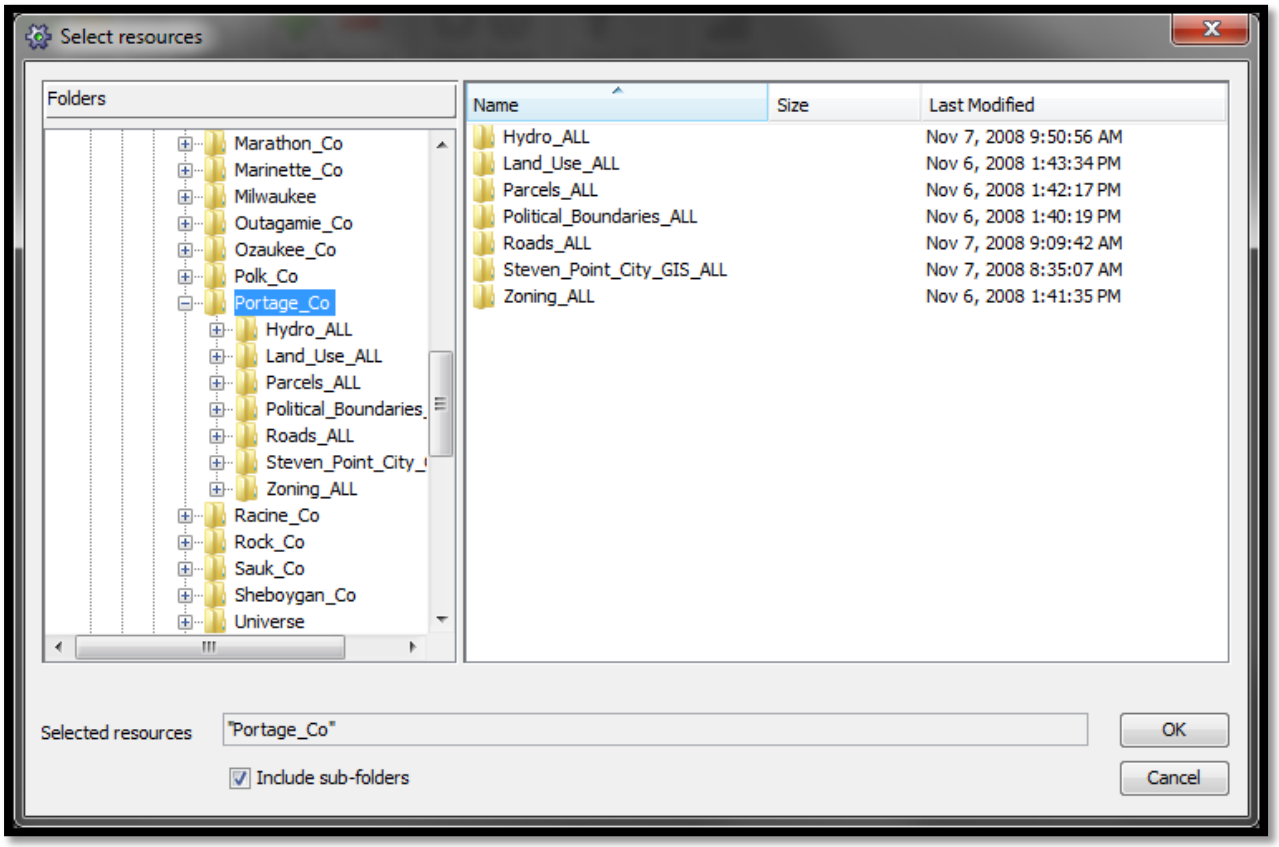
Using DROID

DROID is a java application. After downloading the tool, unzip the files into their "final" location, there is no installer. To run the tool, run the file called droid.bat (Mac and Linux versions will have a droid.sh). Detailed instructions for installation and set-up can be found in the user guide

1. When you run DROID, it will bring you directly to the main interface. A tab called Untitled-1 will appear. Before doing anything, it is important to set up your preferences. Any changes to the preferences will only take effect on profiles created *after* the changes are made.
    a. Go to Tools > Preferences.
    b. Ensure that "Analyse the contents of archive files" is checked if you wish to see files inside of .zip, .tar.gz, or .gzip files.
    c. If you are analyzing file fixity or scanning for duplicate files, generating a hash for each file can be useful. Click the check box and choose *md5*.
    A *hash* is a unique string of letters and numbers based on the actual contents of the file. If two files are identical, they will have identical hash values. If a file changes, so will its hash value. This value is useful for Fixity and for duplicate detection. However, it increases the time required for scanning *significantly*.

d. Under the “Signature Updates” tab, I recommend changing the update frequency to “Every time DROID starts up”, especially if you have file formats that are actively updated.

2. Once you’re satisfied with your preferences, close any automatically created profile and click the green “New” button. This will create a new, untitled profile with the settings specified.
3. Next, add the directory or directories to be analyzed by clicking the Green “Add” button. Be sure to check “Include sub-folders” in the “Select Resources” form. In the example below, the “Portage_Co” directory is the target for profiling.



4. Before starting the scan, save your profile. Click the Save Button. Save the profile somewhere that you can access it. I recommend saving with a date in the filename so that you can compare new profiles to this old one in the future. Saving at this point allows for easily running the analysis again if anything goes wrong.

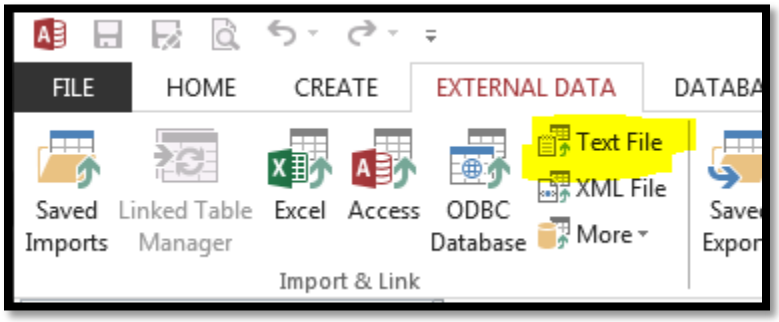
5. Click Start. This could take from 30 seconds to small folders to hours and days for very large folders. If running a large profile in the background, you can use the "throttle" slider at the bottom of the application window to slow down the scan and free up more resources.
   You will be able to start inspecting the profile during the scan.
6. When the scan is completed, I recommend saving immediately. This can take several minutes or longer for large profiles.

7. There are a few built-in tools to analyze the profile. First, you can filter the results by any of the fields included in the report. You can also use the "Report" tool to generate a report.
   a. Click the report button and then select the profile(s) you wish to report on.
   b. Under "Select Report" you have a few options, choose the most appropriate for your analysis. Reports can be as simple as file counts and sizes to comprehensive breakdowns of each individual file, file type, and more.
   c. You can view the report as a stylized document right in the application or click Export to export the report to XML, HTML, TXT, or PDF.

8. The most useful function in DROID is the ability to export the profile to a text file and then bring it into a database for further analysis.
   a. Click Export
   b. Select the profile(s) you wish to export. Make sure that "One row per file" is selected if you want a comprehensive list. If you're only interested in the file formats present, use the other option (You can select the default option for this in the preferences). Click Export Profiles
   c. I recommend saving the export in the same location you save the profile and again using a date in the file name so that you can compare to new reports in the future. By default, this file with not have a file extension, but I recommend adding .csv, because it is a text file of comma separated values.

Importing a DROID export into MS Access

1. Open Microsoft Access and make a new blank desktop database in your desired location. You can reuse the database when you make a new export in the future.
2. In the ribbon, under the External Data tab and in the Import & Link section, click "Text File" to add data from text.



3. In the form that opens, ensure that "Import the source data into a new table in the current database" is selected. Browse for your export CSV and click Open.
4. Click OK in the "Get External Data" form.
5. Choose Delimited and click Next
6. Check the box "First Row Contains Field Names", ensure Comma is the selected delimiter, and that Text Qualifier is set to a double quotation. If the table is looking correct in the preview, click Next.
7. In this view, you are specifying data types for the fields. There are a couple very important things to change here. Use the horizontal scroll bar to find the "ID" column. Change the Data Type to "Integer". Do the same for the "PARENT ID" and "FORMAT COUNT" fields. Do the same for the SIZE field, except choose "Long Integer". Click Next.
8. This view asks about keys. In this case, our data has an ID field already called "ID". Select "Choose my own primary key" and choose "ID". Click Next.
9. The last view will have you name your table. I recommend including a date if you plan to make future DROID profiles and wish to compare them.
10. Click Finish and Save your database.

If all went according to plan, you now have a table listing every file in the scanned directory. The possibilities for what we can do with this data are plentiful.

Analyzing a DROID export in MS Access: Duplicate Detection

There is tons of possibility for analysis in Access, one example is duplicate detection. Duplicate detection can take advantage of some built-in tools in MS Access Data fields for analysis include an ID number (based on the location in the file tree), a Parent ID, File Path, URI (based on file path), the size in bytes, profiled file type information, extension, filename, etc.

See the duplicate detection workflow document for information on duplicate detection using Duplicate Cleaner Free 4.0.5.

Remember that if you add/move/remove files in your directory, IDs and Parent_IDs will change, so if you are looking for a persistent ID, the best bet is the URI field. It's a little long, but a file in the same place will have the same UID even if the file itself is changed or renamed (but not if it's moved).

1. Duplicate Detection
    a. On the ribbon, under create, click "Query Wizard"
    b. On the New Query form, choose "Find Duplicates Query Wizard". Click OK.
    c. Select the table in which you would like to detect dupes.
    d. When the wizard asks "Which fields might contain duplicate information?" select "MD5_HASH". If you don't have this field available, make a new profile paying special attention to step 1c in "*Using Droid*".
    e. When the wizard asks about other fields you wan to include in the query, choose what you think will be useful. I normally use the following:
    ID, PARENT_ID, URI, NAME, SIZE, EXT
    f. If you want to save your query, give it a name. Otherwise, accept the default and click Finish.
    g. Some important notes about duplicate detection:
        i. Duplicates and "originals" or "masters" or whatever are all displayed. If you have three identical files, all three will be shown.
        ii. Duplicates identified may not share a file name. In the duplicate detection ran for this example, DROID found 6 files with identical hash and size, but differing names. Armed with this information, one can now identify which of these files should be preserved and which can be discarded.