





Making Sense of a Collection

DROID tool demo

David Clipsham

6 December 2013

DROID



- File format identification utility
- Scans internal byte sequences of files
- Uses PRONOM registry signature files at its core
- Both command line and GUI interfaces
- Embedded within Tessella SDB and other tools
- <http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm>

PRONOM

- File format registry
- Over 1000 entries (PUIDs)
- Format extensions, mime/media types, links to documentation
- File format identification signatures (for DROID!)
- <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

DROID Results

DROID v6.1.2

File Edit Run Filter Report Tools Help

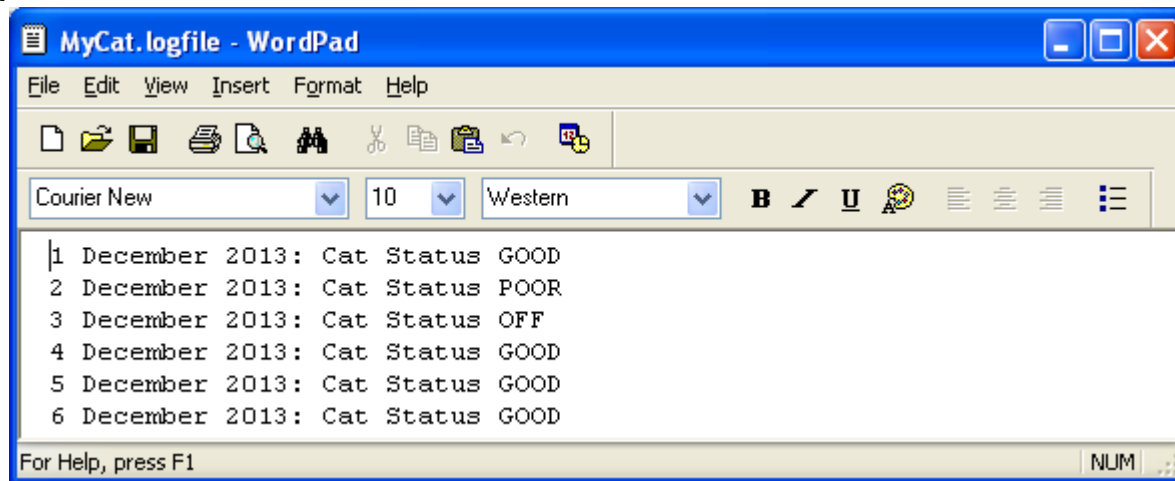
New Open Save Export Add Remove Start Pause Filter On Report

Untitled-1 Untitled-2 x Untitled-3 x Untitled-4 x Untitled-5 x Untitled-6 x Untitled-7 x

Resource	Extension	Size	Last modified	Ids	Format	Version	Mime type	PUID	Method	Hash
Q:\Digital Preservation Team...			05/12/13 16:06							
MyFolder			05/12/13 14:58							
MyOtherDuplicateCat.jpg	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange Format	1.01	image/jpeg	fmt/43	Signature	76e7f18f24b9d7e1ef36932973...
MyZippedCat.zip	zip	23.7 KB	05/12/13 15:02		ZIP Format		application/zip	x-fmt/263	Signature	53e12f52e89653180848af466...
MyCat.jpg	jpg	23.8 KB	05/12/13 14:48		JPEG File Interchange Format	1.01	image/jpeg	fmt/43	Signature	76e7f18f24b9d7e1ef36932973...
My7ZippedCat.7z	7z	23.8 KB	05/12/13 15:03		7Zip format			fmt/484	Signature	8c8655c7ce7de00c7ca4132fec...
MyBrokenCat.jpg	jpg	22.9 KB	05/12/13 14:49							0575a0a40297fd165b3f31ead...
MyCat.jpg	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange Format	1.01	image/jpeg	fmt/43	Signature	76e7f18f24b9d7e1ef36932973...
MyCat.logfile	logfile	201 bytes	05/12/13 16:06							86d9e96c2d441d2c677a7279a...
MyCommaSeparatedCat.csv	csv	24 bytes	05/12/13 14:52		Comma Separated Values		text/csv	x-fmt/18	Extension	d3c53fa9ae8a0f33f1590cd7bd...
MyCorruptedPDFCat.pdf	pdf	15 bytes	05/12/13 14:57							8fca85ab7ebcb01848aa50f6e3...
MyDisguisedCat.jpg	jpg	8 bytes	05/12/13 14:51		Windows New Executable			x-fmt/410	Signature	eeb188cc97b2268822a28854c...
MyDuplicateCat.jpg	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange Format	1.01	image/jpeg	fmt/43	Signature	76e7f18f24b9d7e1ef36932973...
MyEmptyCat.jpg	jpg	0 bytes	05/12/13 14:48							d41d8cd98f00b204e9800998e...
MyExecutableCat.exe	exe	8 bytes	05/12/13 14:51		Windows New Executable			x-fmt/410	Signature	eeb188cc97b2268822a28854c...
MyOtherDuplicateCat.jpg	jpg	23.8 KB	05/12/13 14:47		JPEG File Interchange Format	1.01	image/jpeg	fmt/43	Signature	76e7f18f24b9d7e1ef36932973...
MyPDFCat.pdf	pdf	26.5 KB	05/12/13 14:55		Acrobat PDF 1.5 - Portable Doc...	1.5	application/pdf	fmt/19	Signature	43d1177a3a94246ab99e725f5...
MyPlainTextCat.txt	txt	17 bytes	05/12/13 14:52		Plain Text File		text/plain	x-fmt/111	Extension	564447a3c352895f332c29400...

Unidentified Files

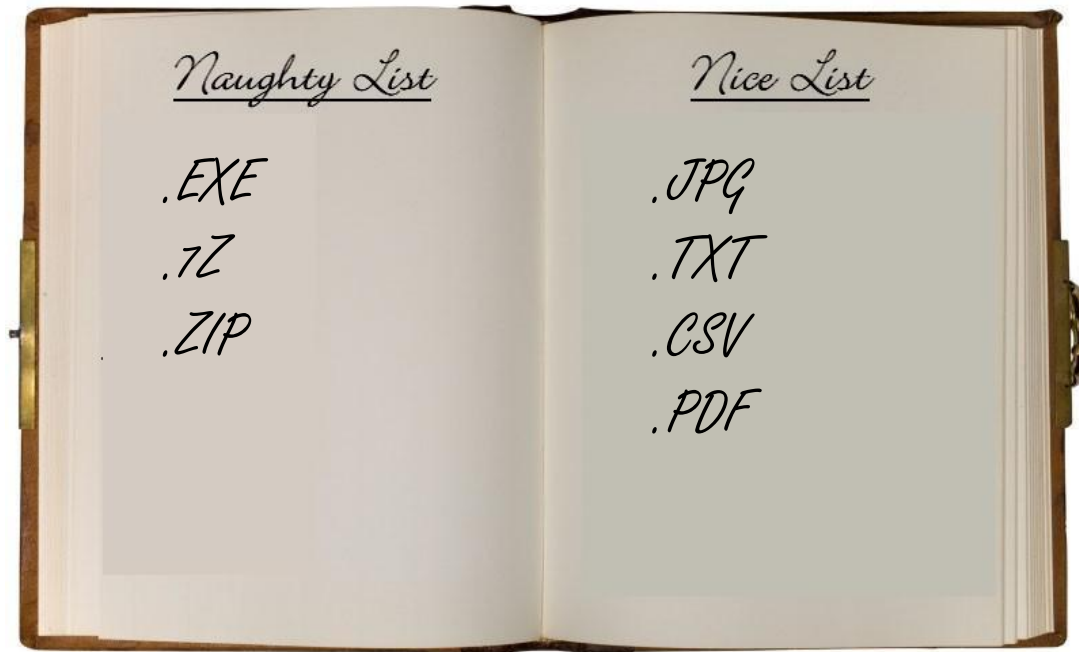
- Unidentified files prompt an intervention – we want to understand why
- Common reasons:
 - Genuine formats PRONOM/DROID simply doesn't have yet (let us know!)
 - System files – logs, configs etc
 - 'Empty' files – i.e. files containing no bytes
 - Corrupted files



- Unidentified files are rare – a recent real world disk drive contained 1277 files, of which 2 were unidentified – both were 'empty' zero-byte files.

White-list/Black-list

- White/black-listing driven by PUIDs






E	M
NAME	MD5_HASH
DJC Collection	
MyCat.jpg	76e7f18f24b9d7e1ef369329739adc84
MyFolder	
MyOtherDuplicateCat.jpg	76e7f18f24b9d7e1ef369329739adc84
MyEmptyCat.jpg	d41d8cd98f00b204e9800998ecf8427e
MyBrokenCat.jpg	0575a0a40297fd165b3f31ead85d6edb
MyExecutableCat.exe	eeb188cc97b2268822a28854c950f1c4
MyCommaSeparatedCat.csv	d3c53fa9ae8a0f33f1590cd7bd4972f5
MyPlainTextCat.txt	564447a3c352895f332c2940026fe410
MyDuplicateCat.jpg	76e7f18f24b9d7e1ef369329739adc84
MyOtherDuplicateCat.jpg	76e7f18f24b9d7e1ef369329739adc84
MyPDFCat.pdf	43d1177a3a94246ab99e725f5485ecf
MyCorruptedPDFCat.pdf	8fca85ab7ebcb01848aa50f6e373cc58
MyZippedCat.zip	53e12f52e89653180848af466da54c1c
MyCat.jpg	76e7f18f24b9d7e1ef369329739adc84
My7ZippedCat.7z	8c8655c7ce7de00c7ca4132fec4a6795

De-duplication

- Built in MD5 checksum generation

Other considerations

 MyDisguisedCat.jpg jpg  8 bytes 05/12/13 14:51  Windows New Executable

- Files with incorrect extension: May not be malicious, but may require understanding
- Archival containers: ZIP, GZIP, TAR can be inspected by DROID. 7ZIP, RAR, ISO and other 'disk image' formats cannot (yet)

NAME	EXTENSION_MISMATCH
DJC Collection	FALSE
MyCat.jpg	FALSE
MyFolder	FALSE
MyOtherDuplicateCat.jpg	FALSE
MyEmptyCat.jpg	FALSE
MyBrokenCat.jpg	FALSE
MyExecutableCat.exe	FALSE
MyCommaSeparatedCat.csv	FALSE
MyPlainTextCat.txt	FALSE
MyDuplicateCat.jpg	FALSE
MyOtherDuplicateCat.jpg	FALSE
MyPDFCat.pdf	FALSE
MyCorruptedPDFCat.pdf	FALSE
MyZippedCat.zip	FALSE
MyCat.jpg	FALSE
My7ZippedCat.7z	FALSE
MyDisguisedCat.jpg	TRUE
MyCat.logfile	FALSE

Tools & Resources

- DROID - <http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm>
- PRONOM - <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>
- DROID – How to Use it and How to Interpret Your Results: <http://www.nationalarchives.gov.uk/documents/information-management/droid-how-to-use-it-and-interpret-results.pdf>

Thank you

From both me and my cat (Squee!)

Any questions?

