Comparing DROID signature files – EOF vs no EOF

An exploration into the usage of the EOF section of a PRONOM format signature, and the ‘fast’ scanning DROID (v6) function

www.natlib.govt.nz

Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY
OF NEW ZEALAND

Document Control

Revision history

Revision	Date	Author	Reason for Change
V1	9.02.2011	J Gattuso	Final

Contents

Comparing DROID signature files – EOF vs no EOF

Summary

This short paper presents the results found when the End Of File (EOF) section is removed from the current Droid Signature (v55).

A second test is completed, in which the 'Maximum Scan Bytes' function is tested with the same signature files.

A recommendation is made to repeat these tests, and extend the scope of the source files.

A recommendation is made to consider the case for removing some EOF markers if there is an efficiency gain in doing so.

A recommendation is made to assess the impact of using the 'Fast' mode of Droid v6.

Method

The current DROID signature (v55) file was downloaded from the DROID resource pool.

A version of the signature v55 XML was created, by manually removing all the EOF references from the <InternalSignatureCollection>

Both signature files (v55 & v55-NO EOF) were used in DROID v6 against the test set of digital objects with the 'Maximum Scan Byte' value set to default (64435 bytes) and the test repeated with this variable set to -1

In all tests container-signature-20110205 was used.

Upon completion, all profiles where exported as CSV files, and imported into MySQL for analysis

Source of digital objects

The test data set used is a collection of objects NLNZ is currently using for a larger series of tests looking into.

There are ~27,000 files, of at least 60 different source PUIDs (depending on the choice of baseline)

Each file is represented in the collection twice, once with its original file extension, once without.

Some source PUIDs only have 2 source files, others have 1000.

Results

Interpreting the results:

The first column is a list of all the distinct PUIDs found in the DROID log files. The 2nd column is a count of total number of times each PUID is offered in the whole log for the vanilla version of signature v55. The 3rd column of is a count of total number of times each PUID is offered in the whole log for the 'NO-EOF' version of signature v55.

MySQL Query

The following MySQL Query was used to extract the data:-

```
SELECT `PUID`,`DROID_V`,`SIG_V`,`SPEED`,

 COUNT(distinct NAME) as `All`

FROM sourcelist, main_v55_only

WHERE main_v55_only.NAME = sourcelist.SourceFileName

GROUP BY `PUID`,`DROID_V`,`SIG_V`,`SPEED` ORDER BY `DROID_V` ASC, `SIG_V` ASC, `SPEED`;
```

The data fields are a direct import of the of the DROID log CSV, with the addition of some labelling columns for DROID version, Signature Version and 'Speed'

main_v55_only is a table containing only DROID 6, signaturev55 results (as used in this test)

sourcelist is a list of all the files used in the test.

Expectations

These tests should be self validating. If there is no change in how a specific file or set of files is being handled by DROID then each of the counts should match. Any delta indicates the file is being characterised differently by a specific set of test parameters

It is worth noting that as some files may get a multiple PUID assertion by DROID, the number of files represented for a specific PUID may exceed the number of 'baselined' files presented to DROID of that specific type.

The baselining of the files used in this particular test is not specifically relevant. The source files should be considered as a fixed set of unknown files, and the results of the various tests should be considered as the attempt of each specific DROID implementation to classify this unknown set of files. Of interest therefore is the consistency of DROID performance, and any delta between test conditions, not the specific performance of an individual file. This data is available as a MySQL database, and can be supplied to interested parties.

Test 1 – 'SLOW' Full results

PUID	v55	v55-NO
fmt/100	8	8
fmt/101	1000	
fmt/11	34	34
fmt/111	12	12
fmt/116	310	310
fmt/117	2	2
fmt/12	56	56
fmt/126	64	64
fmt/132	98	98
fmt/133	2	2
fmt/134	980	1000
fmt/14	2	2
fmt/141		
fmt/142		
fmt/143		
fmt/144	4	
fmt/145	4	
fmt/146	4	
fmt/147	4	
fmt/148	4	
fmt/149	3	
fmt/15	86	86
fmt/157	4	
fmt/158	4	
fmt/16	1020	1020
fmt/17	1038	1040
fmt/18	1232	1236
fmt/189	20	20
fmt/19	1304	1306
fmt/198	4	922
fmt/199	290	290
fmt/20	1268	1268
fmt/214	6	6
fmt/276	254	254
fmt/278	6	6
fmt/279	500	500
fmt/3	12	12
fmt/353	1000	
fmt/354	30	30
fmt/355	12	12
fmt/39	22	22

Comparing DROID signature files – EOF vs no EOF

fmt/4	38	38
fmt/40	1984	1984
fmt/41	934	936
fmt/42	220	220
fmt/43	970	974
fmt/44	997	997
fmt/45	78	78
fmt/46	78	78
fmt/47	78	78
fmt/48	78	78
fmt/49	78	78
fmt/5	130	130
fmt/50	136	136
fmt/51	136	136
fmt/53	186	186
fmt/6	1000	1000
fmt/61	1006	1006
fmt/96	6	6
fmt/99	50	50
x-fmt/111	501	501
x-fmt/135	2	2
x-fmt/219	1000	1000
x-fmt/263	32	32
x-fmt/279	10	10
x-fmt/280		1000
x-fmt/31	500	500
x-fmt/385	4	4
x-fmt/386	4	4
x-fmt/387	1000	2000
x-fmt/388		2000
x-fmt/390	1000	1000
x-fmt/391	1000	1000
x-fmt/394	148	148
x-fmt/398	48	48
x-fmt/399		2000
x-fmt/409	2	2
x-fmt/411	2	2
x-fmt/62	500	500
x-fmt/80	28	28
x-fmt/92	126	126
No PUID Found	1538	1511
Total Assertions	26331	31225

Test 2 – 'FAST' Full results

PUID	v55	v55-NO
fmt/100	8	8
fmt/101	1000	
fmt/11	34	34
fmt/111	12	12
fmt/116	310	310
fmt/117	2	2
fmt/12	56	56
fmt/126	64	64
fmt/132	98	98
fmt/133	2	2
fmt/134	980	1000
fmt/14	2	2
fmt/141	427	427
fmt/142	427	427
fmt/143	427	427
fmt/144	4	
fmt/145	4	
fmt/146	4	
fmt/147	4	
fmt/148	4	
fmt/149	3	
fmt/15	86	86
fmt/157	4	
fmt/158	4	
fmt/16	1020	1020
fmt/17	1038	1040
fmt/18	1254	1258
fmt/189	20	20
fmt/19	1304	1306
fmt/198	3	922
fmt/199	290	290
fmt/20	1274	1274
fmt/214	6	6
fmt/276	254	254
fmt/278		
fmt/279	500	500
fmt/3	12	12
fmt/353	1000	
fmt/354	2	2
fmt/355	12	12

fmt/39	22	22
fmt/4	38	38
fmt/40	1984	1984
fmt/41	934	936
fmt/42	220	220
fmt/43	970	974
fmt/44	997	997
fmt/45	78	78
fmt/46	78	78
fmt/47	78	78
fmt/48	78	78
fmt/49	78	78
fmt/5	130	130
fmt/50	136	136
fmt/51	136	136
fmt/53	186	186
fmt/6	146	146
fmt/61	1006	1006
fmt/96	6	6
fmt/99	50	50
x-fmt/111	501	501
x-fmt/135	2	2
x-fmt/219	1000	1000
x-fmt/263	32	32
x-fmt/279	10	10
x-fmt/280		1000
x-fmt/31	500	500
x-fmt/385	4	4
x-fmt/386	4	4
x-fmt/387	1000	2000
x-fmt/388		2000
x-fmt/390	1000	1000
x-fmt/391	1000	1000
x-fmt/394	148	148
x-fmt/398	48	48
x-fmt/399		2000
x-fmt/409	4	4
x-fmt/411		
x-fmt/62	500	500
x-fmt/80	28	28
x-fmt/92	126	126
No PUID Found	1965	1938
Total Assertions	27178	32073

Delta only – All tests combined:

	FAST	FAST2	SLOW	SLOW2
PUID	**v55**	**v55-NO**	**v55**	**v55-NO**
fmt/101	1000		1000	
fmt/134	980	1000	980	1000
fmt/141	427	427		
fmt/142	427	427		
fmt/143	427	427		
fmt/144	4		4	
fmt/145	4		4	
fmt/146	4		4	
fmt/147	4		4	
fmt/148	4		4	
fmt/149	3		3	
fmt/157	4		4	
fmt/158	4		4	
fmt/17	1038	1040	1038	1040
fmt/18	1254	1258	1232	1236
fmt/19	1304	1306	1304	1306
fmt/198	3	922	4	922
fmt/20	1274	1274	1268	1268
fmt/278			6	6
fmt/353	1000		1000	
fmt/354	2	2	30	30
fmt/41	934	936	934	936
fmt/43	970	974	970	974
fmt/6	146	146	1000	1000
x-fmt/280		1000		1000
x-fmt/388		2000		2000
x-fmt/399		2000		2000
x-fmt/409	4	4	2	2
x-fmt/411			2	2
No PUID Found	1965	1938	1538	1511
Total Assertions	27178	32073	26331	31225

Delta Only – FAST Vanilla vs NO-EOF

	FAST	FAST2
PUID	**v55**	**v55-NO**
fmt/101	1000	
fmt/134	980	1000
fmt/14	2	2
fmt/144	4	
fmt/145	4	
fmt/146	4	
fmt/147	4	
fmt/148	4	
fmt/149	3	
fmt/157	4	
fmt/158	4	
fmt/17	1038	1040
fmt/18	1254	1258
fmt/19	1304	1306
fmt/198	3	922
fmt/278		
fmt/353	1000	
fmt/41	934	936
fmt/43	970	974
fmt/6	146	146
x-fmt/280		1000
x-fmt/388		2000
x-fmt/399		2000
No PUID Found	1965	1938
Total Assertions	10627	14522

Delta Only – SLOW Vanilla vs NO-EOF

	SLOW	SLOW2
PUID	**v55**	**v55-NO**
fmt/101	1000	
fmt/134	980	1000
fmt/144	4	
fmt/145	4	
fmt/146	4	
fmt/147	4	
fmt/148	4	
fmt/149	3	
fmt/157	4	
fmt/158	4	
fmt/17	1038	1040
fmt/18	1232	1236
fmt/19	1304	1306
fmt/198	4	922
fmt/20	1268	1268
fmt/278	6	6
fmt/353	1000	
fmt/354	30	30
fmt/41	934	936
fmt/43	970	974
x-fmt/280		1000
x-fmt/388		2000
x-fmt/399		2000
No PUID Found	1538	1511
Total Assertions	11335	15229

Delta Only - FAST vs SLOW (Vanilla)

	FAST	SLOW
PUID	**v55**	**v55**
fmt/141	427	
fmt/142	427	
fmt/143	427	
fmt/18	1254	1232
fmt/6	146	1000
fmt/198	3	4
fmt/20	1274	1268
fmt/278		6
x-fmt/409	4	2
x-fmt/411		2
fmt/354	2	30
No PUID Found	1965	1538
Total Assertions	5929	5082

Delta Only - FAST vs SLOW (NO-EOF)

	FAST	SLOW
PUID	**v55-NO**	**v55-NO**
fmt/141	427	
fmt/142	427	
fmt/143	427	
fmt/18	1258	1236
fmt/20	1274	1268
fmt/278		6
fmt/354	2	30
fmt/6	146	1000
x-fmt/409	4	2
No PUID Found	1938	1511
Total Assertions	5903	5053

Conclusions

It should be noted that for a large number of tested objects the EOF aspect of the DROID signature had no impact on its consistent PUID assertion (i.e. all variations offered the same total number of files of a specific PUID)

These results have only been analysed in a coarse way to establish basic trend observations.

To accurately understand the changes in PUID assignation it would be required to undertake a deeper study of this data set.

Primary findings are that:

(1) Some file types are not affected by removal of EOF patterns and by DROID 'fast' mode

(2) Some files types behave inconsistently when the EOF section is removed from the signature

(3) Some files types behave inconsistently when 'Maximum Byte Scan size' feature is used in DROID v6

Recommendations

It is recommend that:

(1) other institutions recreate this experiment to valid these trends – if similar trends are observed the relevance of the EOF section should reconsidered for some PUIDS if there is found to be an efficiency gain for removing this element

(2) a deeper study be undertaken with a ground-truthed source file set that covers more format types

(3) DROID users consider the impact of running DROID in 'fast' mode, and how this feature demonstrably affects PUID assignation for some formats.

   a. This test should be extended to establish if there is a 'safe' threshold that results in consistent results in both 'slow' and 'fast' modes.

   b. The inconsistency may be observed as either 'no PUID found' or 'alternate PUID(s) found'. This notion should be explored in more detail, specifically to consider the impact of the two types of inconstancy on any operational workflow.