

## Assessment of file format testing tools

This report is an assessment of available file format testing tools, and an assessment of whether or how they could be used as part of a Digital Asset Assessment Tool (DAAT).

Assessment of file format testing tools .....	1
<b>1. Introduction.....</b>	<b>2</b>
<b>2. DROID summary assessment .....</b>	<b>4</b>
2.1 Overview .....	4
2.2 Range of the test sample.....	4
2.3 Results: Positive ID.....	4
2.4 Results: Tentative ID .....	4
2.5 Results: Error.....	5
2.6 Results: Not identified.....	5
2.7 Conclusions .....	5
<b>3. JHOVE summary assessment.....</b>	<b>7</b>
3.1 Range of the test sample.....	7
3.2 Identification results .....	7
3.3 Conclusions .....	8
<b>4. AIHT / Empirical Walker summary assessment .....</b>	<b>9</b>
4.1 Overview .....	9
4.2 Range of test sample.....	9
4.3 Some aspects of Empirical Walker tool .....	9
4.4 Scoring with Empirical Walker .....	10
4.5 Conclusions .....	10
<b>5. Using DROID, JHOVE or Empirical Walker as part of a D-PAS tool.....</b>	<b>11</b>
5.1 Overview .....	11
5.2 DROID integration with D-PAS .....	11
5.3 JHOVE integration with D-PAS .....	12
5.4 Empirical Walker integration with D-PAS .....	12

APPENDIX 1: DROID test results

APPENDIX 2: Sample outputs from JHOVE test results

## 1. Introduction

Understanding file formats is an important part of preserving digital assets. The Digital Asset Assessment Tool (DAAT) project aimed to assess three available tools and evaluate their potential usefulness as part of the Digital Preservation Assessment Survey (D-PAS) tool. Each tool offered some form of automated file format identification.

The three tools evaluated were:

**1. DROID** (Digital Record Object Identification), a software tool developed by The National Archives to perform automated batch identification of file formats. It is the first in a planned series of tools developed by The National Archives under the umbrella of its PRONOM<sup>1</sup> technical registry service.  
(<http://droid.sourceforge.net/wiki/index.php/Introduction>)

**2. JHOVE** - JSTOR/Harvard Object Validation Environment. JSTOR<sup>2</sup> and the Harvard University Library are collaborating on a project to develop an extensible framework for format validation. JHOVE provides functions to perform format-specific identification, validation, and characterization of digital objects.  
(<http://hul.harvard.edu/jhove/>)

**3. AIHT<sup>3</sup> at Stanford University:** Automated Preservation Assessment of Heterogeneous Digital Collections, and the **Empirical Walker** tool.  
(<http://www.dlib.org/dlib/december05/johnson/12johnson.html>)

Within the DAAT project, our aim was not to make a general report on whether DROID or JHOVE work at all, but whether they have useful qualities which will help us identify the preservation needs of our assets.

The following outline shows the scope of these tests:

**Object:** to assess the usefulness and viability of the file format identification tools DROID, JHOVE and Empirical Walker, within the context of the DAAT project.

**Outcome:** produce a short but detailed technical report on findings, with an intended audience of potential DAAT users. The report should be like a 'consumer' report, with practical guidance and easy-to-understand information.

**Output:**

- A list of all file formats used in the test-bed environments.
- Clearly stated reasons for inclusion of formats in the tests, especially certain formats which appear to be no more than system files.
- Inclusion of some 'stress-test' formats, such as many different types of text files (plaintext, Unicode, MS-DOS), and observation of results.
- Inclusion of some complex objects, such as a Word file with an embedded video, and see if any of the tools manage to identify what's going on in a way that's useful to us.

**Assumptions made:**

- The use of commonly known and supported file formats will affect the preservation of an asset – so D-PAS needs to know about formats.
- We need to identify file formats, and validate them too.
- We need some form of 'deep file' characterisation, to identify the significant characteristics of a format. Metadata extraction is a significant part of this, but not the only dimension of the process.
- In addition to identifying the file format, we may also need to know the software type used to create it, access it, and preserve it. This isn't always necessary, though – a

---

<sup>1</sup> PRONOM is an on-line information system about data file formats and their supporting software products. See <http://www.nationalarchives.gov.uk/pronom/>.

<sup>2</sup> JSTOR: Journal Storage, The Scholarly Journal Archive. See <http://www.jstor.org/>.

<sup>3</sup> AIHT: The Archive Ingest and Handling Test, a project devised and funded by the Library of Congress.

huge number of tools are capable of creating and accessing plain text files, for instance and it's rarely helpful to know exactly which one was used. In fact, many different tools can be used to create a particular digital object.

- Ideally, we'd like to see each digital asset in the context of the information environment that supports it – for example pathnames, dates of access, dates of modification, creator names, etc.
- Other things the D-PAS tool may need to consider are tools for virus checking and fixity checking, but probably not in the context of this task.

## 2. DROID summary assessment

### 2.1 Overview

The DAAT project team downloaded a copy of DROID and made some experiments with it on our own drives.

- 77 files were tested in a test environment.
- 29 files were identified by DROID as 'Positive'.
- 11 files were identified by DROID as 'Tentative'.
- 3 files were reported by DROID as 'Error'.
- 34 files (nearly 50% of the sample) came back as 'Not identified'.

### 2.2 Range of the test sample

Besides familiar formats like text and image format, the sample included a number of system files; files that a digital asset might depend on; complex compound files; and some files created by Mina Creathorn designed specifically to stress-test DROID and the level of detail it could generate.

### 2.3 Results: Positive ID

For common MS Office and other popular formats, DROID successfully identified (for example) .PPT as a Powerpoint Presentation, and .PDF as a Portable Document Format.

However, .XLS files were not identified as Excel spreadsheets, but as 'Binary Interchange File Format (BIFF) Worksheet'. This is a technically correct description of the file format (as opposed to the application which created it), as PRONOM's own information on the format makes clear. It can be confusing to those unfamiliar with these distinctions, however. Another .PPT file was identified as an 'OLE2 Compound Document Format', almost certainly because it contained other media. This is, again, a correct description of the format, but it is less than helpful for assessment or preservation purposes. OLE2 Compound Documents can contain many different types of entity, some of which present much greater management challenges than others.

An .RTF file was positively identified as 'Rich Text Format', but 7 'versions' of the format were offered by DROID as possibilities, with no indication as to which version was more likely to be correct. (For RTF files which don't exploit features specific to a given version, all of them *are* correct, of course. But this is also confusing to the uninitiate.) However, for two .DOC files DROID successfully identified that one was created in MS Word Version 6.0/95 and the other in V2.

For our two 'complex' files, the results were interesting. DROID reported a Word file embedded with an .ASF file as an 'OLE2 Compound Document Format', which presents us with the same problem of unhelpful accuracy which we saw with the powerpoint presentation above. A Word file with a .CSV file embedded fared slightly better, similarly identified as a compound while its csv component was reported as a BIFF worksheet.

Of system files, .EXE, .AX, .SYS, .TLB and .DLL files were identified simply as 'Windows Portable Executable' formats. .BAK and .ODC. files were identified as 'Hypertext Markup Language'.

### 2.4 Results: Tentative ID

In all 'Tentative' cases, DROID identified a named format but nothing else (no version, no PRONOM ID). Two [different?] .CSV files produced the same result – 'Comma Separated Values'.

The results on three text files are interesting in that DROID could not identify any difference between a plain text file, a Unicode text file, or an MS-DOS text file. In all cases, the same nine 'options' were suggested, all apparently equally valid. This is clearly an area where DROID is capable of improvement, since relatively simple techniques are available to

distinguish between these types of text file. It's possible, however, that the distinctions are not, at present, of great relevance to DROID's creators.

## **2.5 Results: Error**

The 3 'Error' files were all zero-length files. The reader's first reaction may be that we were perverse in even asking DROID to look at such files, which don't exactly present great preservation challenges. But zero-length files, though uninteresting in themselves, are meaningful parts of larger file collections and require preserving. It's unhelpful when automated tools report them as errors, particularly if there is no way to shut that error off (we haven't checked if this is, in fact, possible with DROID.)

## **2.6 Results: Not identified**

As noted above, 50% of the sample was 'Not identified' by DROID. The files which DROID fails to recognise include many system files. Yet these files may be needed to access and preserve the digital asset. For example:

.INF files provide the ability to create customized software installation instructions, which include registry entries and destination directories. By pointing to the URLs of files to download, an INF file provides instructions that Internet Explorer uses to install software components.

.HLP files may be needed as an option within the asset.

.DLL Dynamically Linked Library format. A library which is linked to application programs when they are loaded or run. All assets would be dependant on these format files, especially audio/video ones. DROID was 'Positive' about this, but it simply describes it as a 'Windows Portable Executable' format.

## **2.7 Conclusions**

- DROID has information on a large range of formats. It uses internal and external signatures to identify and report the specific file format versions of digital files.
- The behaviour of DROID suggests it only reports on file formats, and nothing else. The file extension isn't everything (and can be misleading) and .PPT doesn't tell a user if it's Powerpoint version 2 or Powerpoint version 9, for instance. DROID sometimes is able to make that distinction for us (when the file is self-describing, as many MS Office binary formats are) but not always – as the RTF example demonstrates.
- It can't yet make distinctions between variations of text formats
- It may not tell us anything definite about the software used to create or access the asset (which is PRONOM's job.)
- The results from DROID can be output in XML, CSV or printer-friendly formats.

DROID is basing its analysis on the most up-to-date information from PRONOM. The tool arrives pre-loaded with format information (which you need to update periodically - you're always prompted to download the latest signature file when you launch the tool), which it then relays in the form of a static table.

The other feature it provides is a traffic light diagnostic; for example, you get a green light if DROID makes a positive ID of your format. But a red light, for example, isn't an indication of whether the asset is at risk. The traffic lights merely indicate that DROID is working.

Further, DROID may not provide any extra information that we can't already obtain by simply looking at the file format extension or a directory listing. DROID isn't really looking at the entire information environment (which D-PAS will need to).

See Appendix 1 for a detailed table of DROID test results.



### 3. JHOVE summary assessment

Soon after we began looking at JHOVE, the Digital Curation Centre published a case study on JHOVE<sup>4</sup>. This prompted the question whether the DAAT project needed to test JHOVE at all, as there is already a lot of available information on its behaviour. However, it was agreed we need to generate evidence of some form of experimentation, and even if the results tell us what we already know, it's still original evidence.

We knew in advance the behaviour of JHOVE was going to be limited (in its test state, it only works on a dozen file formats anyway), but even so we deemed it useful to run it on test file formats held in ULCC systems, and make a note of observations of its behaviour, particularly in the context of the file's importance as an asset.

#### 3.1 Range of the test sample

JHOVE is currently capable of analysing the following file types: AIFF, ASCII, BYSTREAM, GIF, HTML, JPEG, JPEG2000, PDF, TIFF, UTF8, WAVE and XML. We were therefore limited as to the range of files we could test. The following 12 test file formats were subjected to JHOVE analysis:

- .JPEG
- .HTML
- .TBI
- .DLL
- .INC
- .TIFF
- .XML
- .PDF
- .WAV
- .TXT
- .GIF
- .XLS

The TBI and INC files were selected as examples of ASCII format type. The DLL file was selected as an example of BYTESTREAM format type.

#### 3.2 Identification results

##### Result: Well-Formed and Valid

Eight of the test files had their formats correctly identified by JHOVE and marked as Well-formed and valid. The main function JHOVE performs is to validate the format. The DCC report puts it like this: "JHOVE reports validation at two levels: (i) wellformed; and (ii) valid. An object is considered well-formed if all of the individual component structures are correct; in other words, wellformedness is a local property. An object is considered valid if there is overall consistency between the individual component structures / semantic-level requirements; in other words, validity is a global property."

JHOVE also determines format validation conformance with a third characteristic, ie consistency. The JHOVE tutorial states "An object is *consistent* if it is valid and its internally extracted representation information is consistent with externally supplied representation information."

##### Result: Not well-formed

Three of the test files were reported as 'Not well-formed':

The HTML file failed and received an ErrorMessage. This was a Lexical error, caused by bad HTML.

The WAV file also received an ErrorMessage stating 'unexpected end of file'.

With the TXT file, JHOVE encountered an unexpected UTF-16 little-endian encoding in a UTF-8 file.

---

<sup>4</sup> Digital Curation Centre Case Studies and Interviews: JHOVE. Martin Donnelly, HATII, University of Glasgow. March 2006. ISSN 1749-8767.

### 3.3 Conclusions

- Unlike DROID, JHOVE works harder to validate that the file is what it claims to be, and picks up things which could present preservation or rendering problems many years down the line.
- Unlike DROID, it extracts metadata, including properties of the file such as:
  - Date of modification
  - Size
  - Format and version
  - Location of the asset (in RepresentationInformation)
  - MIMEType
- For image formats like TIFF and JPEG, JHOVE also reports on the image metadata itself, whose integrity could be important. It also reports on embedded metadata in a PDF.
- JHOVE also creates a lot of technical information in its outputs whose meaning is not immediately clear to anyone except perhaps a qualified systems expert. There may be further explanations provided in the JHOVE literature, but it was not within our scope to do an exhaustive study of all these things, so we don't know why for example the analysis of the XML file created four pages of code which JHOVE simply calls 'CharacterReferences'.
- JHOVE does not appear to identify software. Although it identifies version, for example PDF Version 1.4, this refers to a format version rather than a version of software used to create or access the asset.
- JHOVE's major limitation is that at the moment it can only analyse a small number of formats. How does it behave when asked to analyse a format for which it's not configured? An MS Excel spreadsheet was included in our test sample; JHOVE reported it as a BYTESTREAM.
- JHOVE can also do an automated crawl of a drive (however we weren't able to download the version in DOS that does this).

The results of our tests indicate that even in a small sample of objects, the tool's success rate is very high. The three 'Not well-formed' results indicate that JHOVE is doing its job, i.e. identifying format weaknesses in particular files, thus highlighting areas which may require preservation action. It may be instructive to compare this with DROID, where a 'Not identified' result doesn't indicate there's anything wrong with the actual file format; rather, it indicates that DROID has failed to identify it.

Clearly JHOVE is going to be essential in a digital preservation context (particularly one which manages to implement the PREMIS<sup>5</sup> model), as it can be used continually to check and recheck each digital object stored in the repository, and by a process of ongoing validation will give you some clues as to whether you're doing something in your preservation actions which might affect the validity of the asset.

See Appendix 2 for a selection of JHOVE test results.

---

<sup>5</sup> PREMIS: PREservation Metadata: Implementation Strategies. For reports of the PREMIS working group, see <http://www.oclc.org/research/projects/pmwg/>.

## 4. AIHT / Empirical Walker summary assessment

### 4.1 Overview

This is a project based at Stanford University in the USA, which the DAAT project learned about in Washington from Keith Johnson (in May 2006). An article about the project was published in *D-Lib magazine* December 2005<sup>6</sup>. The AIHT (Archive Ingest and Handling Test) is a Library of Congress project to create a small real-world digital archive, of which Stanford are one of the four participating institutions.

Unlike DROID and JHOVE, this is not a separate tool as much as an entire methodology, a process for automated assessment of preservation risk. However, the Stanford team *have* evolved an automation tool which is an integral part of their process; they call this tool 'Empirical Walker'.

The D-Lib article describes a fundamentally different approach to that of the National Preservation Office's L-PAS model. Stanford based their model on workflow, not on static collections; they devised software to analyse large numbers of file formats in a hard drive; and created an XML output from the process.

They keep stressing in their report that, for best results, analysis of files and metadata extraction had to be automated, not performed manually. They recognised that automation would allow them to scale the methodology, provide trustworthy data on the collections, maintain control of workflow, and treat a collection as a manageable set of objects.

The results of the analysis take place within a tight framework of pre-determined information, concerning the preservation policy of the organisation, and information about sets of preferred preservation formats.<sup>7</sup> It was tied in an earlier questionnaire form (built by their Technology Assessment Group) which enabled them to evaluate digital objects by classes.

Stanford also made provision for assessing things like context, meaning, value and importance of assets – through questions directed at the content owner.

So their approach was to build two parallel devices:

- 1) A mechanism to raise questions about the nature of the objects, based on classes
- 2) A tool to gather information about an individual object (or related objects)

At time of writing, ULCC do not have a copy of the Empirical Walker software to assess, so this assessment is based entirely on the report given by Stanford staff in their D-Lib article.

### 4.2 Range of test sample

Not known for certain, but Table 1 in the D-Lib article indicates a wide range of formats: Text formats, including ASCII and UTF-8; Image / Graphic formats, including JPEG, GIF, TIFF, BMP, PNG, and Photoshop; MS Office documents, including .DOC, .XLS and .PPT; Audio formats (including WAVE, MP3, and AIFF); Video formats, including MPEG, Real, QuickTime and Windows Media); and possibly more.

### 4.3 Some aspects of Empirical Walker tool

It's an automated workflow; it surveys a collection automatically.

It describes and assesses a collection of digital objects (and judging by published results, can process large numbers of them very quickly).

It drives external assessment tools (including JHOVE, which was used as part the Stanford project).

---

<sup>6</sup> 'The AIHT at Stanford University', by Richard Anderson et al. *D-Lib Magazine* December 2005, Volume 11, Number 12. ISSN 1082-9873.

<sup>7</sup> The UK Data Archive also has a table of preferred 'Principal Ingest Formats' and 'Preservation Formats' for the types of data they handle. (See *Assessment of UKDA and TNA compliance with OAIS and METS standards*, p 89).

It recursively traverses a file directory and:

- Associates external metadata with files
- Calculates checksums automatically
- Identifies the file format
- Analyses the file content
- Discovers relationships between files
- Builds a METS<sup>8</sup> picture of the collection
- Creates structural metadata for the collection

The tool then goes on to create five types of 'Empirical' metadata – presumably, this means metadata associated with each individual digital object it finds. These types include things like fixity and MIME-types. It also uses JHOVE functionality, for example to perform validation.

#### **4.4 Scoring with Empirical Walker**

To get results from the tool, you need to load it with the pre-determined information about the organisation and collection, as mentioned above. This includes information derived from the preservation policy; and a format-scoring matrix, which was used to help determine preferred file formats.

The tool takes all the empirical metadata it has gathered from the directory, and compares it with the pre-loaded information, which enables it to give a score to each individual file.

This is probably based on a comparison of 'desired best practice' for files and formats with a 'snapshot' of what's actually in the drives. From a sophisticated comparison and calculation process, it arrives at a tiered scorecard which, when interpreted correctly, enables the organisation to identify its preservation needs.

Moreover, these results would help identify levels of service appropriate to certain classes of objects, because Stanford's extant evaluations had showed them that not all digital objects are the same, and some of them may require a level of service over and above simple bit preservation.

#### **4.5 Conclusions**

Empirical Walker, if we can believe the claims made for it, seems to be a far superior tool to DROID or JHOVE, even though it uses JHOVE for some of its results. Our reasons for assuming this superiority:

- It does more things – like calculating checksums and analysing file content
- It understands the whole structure of a collection, and relations between files, are important too (this would appear to be a massive improvement on DROID, which works exclusively on a file-by-file basis, regardless of context or location)
- It extracts / generates more metadata than either DROID or JHOVE

The Stanford methodology may also be superior to D-PAS, in that:

- It assumes that there is a workflow, not a static collection of digital objects
- It integrates the automation of the assessment process with questions about the organisation and institutional policies, and does this more successfully than D-PAS
- It delivers scored results based on the analysis of thousands of objects (not 400), compared against fixed matrix systems, which in turn are derived from hard facts about the nature of preferred file formats and local preservation strategies

---

<sup>8</sup> METS: Metadata Encoding and Transmission Standard. See <http://www.loc.gov/standards/mets/>.

## **5. Using DROID, JHOVE or Empirical Walker as part of a D-PAS tool**

### **5.1 Overview**

In theory, a D-PAS tool might be able to integrate with any of these file format analysis tools. If our project's hierarchy model proves workable, then a file format analysis tool could integrate its results at the level we have determined as 'item' level. In particular, the way the D-PAS tool is currently built, a tool such as DROID or JHOVE would come into play at our 'Level 5', and its results would be applicable to Sections 8-11 of the questionnaire.

This assumes that there are two parallel surveys going on: the D-PAS tool collecting the top-level profile information about the organisation and its collections, with the file format tool gathering profile information on individual assets. At some point, the results from a top-down survey and bottom-up survey would have to find some connectivity to make such an exercise work.

There are some fields in D-PAS Sections 8-11 where an overlap, if not an exact correlation, with DROID / JHOVE type information can be clearly seen. For example, the following D-PAS questions from Section 8:

- 'File format type and version'
- 'Application (and version) used to create the asset'
- 'Location metadata'

There remains a question as to how to automate the integration of the tools; whether results from DROID or JHOVE can be extracted and automatically imported into a D-PAS database, and have specific scores assigned to the results, thus adding to the final D-PAS score.

The specific D-PAS questions on in Section 10 (File Formats) and Section 11 (Software) are slightly more complex. They are asking questions about the stability of formats, and the reliability of software, but phrased in such ways that DROID or JHOVE can't really answer. If this integrated approach were to succeed, D-PAS would need a certain amount of reworking.

### **5.2 DROID integration with D-PAS**

In terms of adding an automated 'crawl and assess' feature to a D-PAS tool, we think DROID leaves a lot to be desired. DROID will do an automated crawl of all file formats in a drive, but it will only provide 'static' information on its format, based on whatever information is currently stored in PRONOM. Importantly, DROID isn't really looking 'inside' a file, just reporting on the extension. To put it bluntly, for all your files which end in .TXT, DROID will tell you exactly the same thing for all of them.

In ULCC's test, the sample of files included 'non asset' files with the extensions of, for example, HLP, BAK, INI, EXE. The reason for doing this is that these files may be used to access and read the digital data. The hardware environment also may be dependent on these files. If these files are missing or not safe, the asset is at risk. Unfortunately, DROID does not even recognise any of these system file types. Wherever possible we have commented on why we think they ought to be included in any tool used for risk assessment.

DROID does not extract metadata from a file. This lack of metadata extraction may not be a problem for D-PAS. Metadata is useful overall for preservation, but it's not particularly germane to the risk assessment that D-PAS is meant to do. However, it is relevant to a risk assessment if there is no metadata available at all.

In terms of output, DROID is capable of generating a .CSV file, so the results of a DROID survey could conceivably be integrated with a D-PAS-type database.

### **5.3 JHOVE integration with D-PAS**

Again, it seems feasible to run JHOVE on a collection area identified by the D-PAS survey, and use the results to contribute to the overall D-PAS score.

JHOVE will provide more metadata than DROID, and again some of its fields could be used to identify general or particular file format risks. For example, even simply counting the number of instances of 'Not well-formed' or 'ErrorMessage' generated by a JHOVE sweep could provide a useful statistic. This could feasibly be scored in D-PAS.

The JHOVE field 'RepresentationInformation' could also be used in D-PAS to help relocate a file or an asset for further testing, although this is not a scoring element.

The 'LastModified' field in JHOVE, which is a date field, was considered as something that might add value to a D-PAS survey. It was decided that this date may not actually be a meaningful property of the file; it could equally represent the same value as Created Date.

In short JHOVE does not actually provide any real 'risk' metadata, but it does provide some fields which we could interpret in such ways as to make them contribute towards a risk assessment survey like D-PAS.

We think export formats are possible. The output appears to be structured but we're not sure if it is. If we can export, we'd like to find ways to feed the data from certain fields into the D-PAS database and get them to score.

### **5.4 Empirical Walker integration with D-PAS**

As described above, the Stanford solution is an integrated package of which Empirical Walker is one component. There may be a way to separate out the Empirical Walker tool component and include it as a step within D-PAS. However:

Empirical Walker is so completely integrated into the Stanford methodology that there seems to be little point in separating it out.

The Stanford methodology is fundamentally different to D-PAS, meaning the results from Empirical Walker may not be compatible with D-PAS.

## File Format Assessment: DROID test results

DROID test My findings so far 01-Aug-2006 Mina Creathorn		My list includes "non asset" files with the extensions of, for example, .hlp, bak, ini, exe. My reasons for including these - these files may be used to access and read the digital data. The hardware environment also may be dependent on these files. If these files are not safe or missing the asset is at risk. DROID does not recognise many file types and wherever possible I have commented on why I think they ought to be included in any tool used for risk assessment. For that matter why not assess whole systems and platforms?		
DROID Version	V1.0			SigFile Version

## APPENDIX 1

Status	File	Comments	Warning	PUID
Not identified	C:\DROID\\$ncsp\$.inf	INF files provide the ability to create customized software installation instructions, which include registry entries and destination directories. By pointing to the URLs of files to download, an INF file provides instructions that Internet Explorer uses to install your software components.		
Not identified	C:\DROID\WinMgmt.CFG	Configuration file.		
Positive	C:\DROID\+New SQL Server Connection.odc	This connection file may be needed for example to access an sql server database.		fmt/96
Not identified	C:\DROID\1.CAT			
Positive	C:\DROID\1033.MST			fmt/111
Positive	C:\DROID\10500107			fmt/111

Status	File	Comments	Warning	PUID
Positive	C:\DROID\6to4svc.dll	Dynamically Linked Library format. A library which is linked to application programs when they are loaded or run. All assets would be dependant on these format format files, especially audio/video ones.		null
Not identified	C:\DROID\85S874.FO_			
Not identified	C:\DROID\access.hlp	This file may be needed as an option within the asset.		
Positive	C:\DROID\acelpdec.ax			null
Not identified	C:\DROID\aclui.chm			
Positive	C:\DROID\acpi.sys			null
Positive	C:\DROID\activeds.tlb			null
Positive	C:\DROID\act_rs.png			fmt/11
Not identified	C:\DROID\adcjavas.inc			
Not identified	C:\DROID\Address Book.lnk			
Not identified	C:\DROID\ADMEXS.DL_			
Not identified	C:\DROID\ADMTOOLW.CH_			
Not identified	C:\DROID\adojavas.inc			
Not identified	C:\DROID\ADSUTIL.VB_			
Positive	C:\DROID\agentsvr.exe			null
Not identified	C:\DROID\AGT0404.DL_			
Tentative	C:\DROID\amipro.sam			null
Not identified	C:\DROID\appstar3.ani			
Positive	C:\DROID\Argentinien2001.PPT			fmt/126
Not identified	C:\DROID\ARIAL.TT_			
Not identified	C:\DROID\ATOMIC.WM_			
Not identified	C:\DROID\Autoscript.chs			

Status	File	Comments	Warning	PUID
Positive	C:\DROID\excel.xls			fmt/59
				fmt/60
				fmt/111
Positive	C:\DROID\excel4.xls			fmt/57
Not identified	C:\DROID\FOLDER.ICO			
Positive	C:\DROID\FreeTaxGuide.pdf			fmt/17
Not identified	C:\DROID\glob.settings.js			
Positive	C:\DROID\in the garden.JPG			fmt/43
Error	C:\DROID\interrupted.lock		Zero-length file	
Positive	C:\DROID\londontoexeter.htm			fmt/96
Tentative	C:\DROID\lotus.wk4			null
				null
Positive	C:\DROID\MSDE2000A.exe			null
Error	C:\DROID\MSDOS.SYS		Zero-length file	
Tentative	C:\DROID\msdos.txt	I created an MS-DOS text file but DROID did not identify it explicitly.		null
				null
				null
				null

Status	File	Comments	Warning	PUID
Positive	C:\DROID\P7300969.JPG			fmt/41
Positive	C:\DROID\powerpnt.ppt			fmt/111
Tentative	C:\DROID\presenta.shw			null
Positive	C:\DROID\projects-export.rm.xml			fmt/101
Not identified	C:\DROID\project_outputs.tbi			
Not identified	C:\DROID\PUTTY.RND			
Not identified	C:\DROID\quattro.wb2			
Positive	C:\DROID\richrtf.rtf			fmt/45
				fmt/46
				fmt/47
				fmt/48
				fmt/49
				fmt/50
				fmt/51
Positive	C:\DROID\Seal Petition.pdf			fmt/18

Status	File	Comments	Warning	PUID
				null
				null
				null
				null
				null
				null
				null
				null
				null
Positive	C:\DROID\Un verdadero hijo de puta.asf			fmt/132
Tentative	C:\DROID\unicodetxt.txt	DROID did not identify this as a Unicode text file.		null
				null
				null
				null
				null
				null
				null
				null
				null
				null
Positive	C:\DROID\winword.doc			fmt/39
Positive	C:\DROID\winword2.doc			fmt/38
Tentative	C:\DROID\wordpfct.wpd			null

Status	File	Comments	Warning	PUID
			fmt/62	Binary Interchange File Format (BIFF) Workbook
			fmt/111	OLE2 Compound Document Format

## File Format Assessment

## APPENDIX 2

Sample outputs from ULCC's test of JHOVE, August 2006

```
JhoveView (Rel. 1.0, 2005-05-26)
Date: 2006-08-08 14:45:31 BST
RepresentationInformation: C:\DROID\in the garden.JPG
ReportingModule: JPEG-hul, Rel. 1.1 (2004-12-10)
LastModified: 2006-06-07 10:48:53 BST
Size: 29164
Format: JPEG
Version: 1.01
Status: Well-Formed and valid
SignatureMatches:
  JPEG-hul
InfoMessage: Unknown TIFF IFD tag: 41985
  Offset: 126
InfoMessage: Unknown TIFF IFD tag: 41986
  Offset: 138
InfoMessage: Unknown TIFF IFD tag: 41987
  Offset: 150
InfoMessage: Unknown TIFF IFD tag: 41988
  Offset: 342
InfoMessage: Unknown TIFF IFD tag: 41989
  Offset: 174
InfoMessage: Unknown TIFF IFD tag: 41990
  Offset: 186
InfoMessage: Unknown TIFF IFD tag: 41992
  Offset: 198
InfoMessage: Unknown TIFF IFD tag: 41993
  Offset: 210
InfoMessage: Unknown TIFF IFD tag: 41994
  Offset: 222
InfoMessage: Unknown TIFF IFD tag: 50341
  Offset: 350
MIMEtype: image/jpeg
Profile: JFIF
JPEGMetadata:
  CompressionType: Huffman coding, Baseline DCT
  Images:
    Number: 1
    Image:
      NisoImageMetadata:
        MIMEType: image/jpeg
        ByteOrder: big-endian
        CompressionScheme: JPEG
        ColorSpace: YCbCr
        ScannerManufacturer: Asahi Optical Co.,Ltd.
        ScannerModelName: PENTAX Optio430RS
        SamplingFrequencyUnit: inch
        ImageWidth: 360
        ImageLength: 480
        BitsPerSample: 8, 8, 8
        SamplesPerPixel: 3
      Scans: 1
      QuantizationTables:
        QuantizationTable:
          Precision: 8-bit
          DestinationIdentifier: 0
```

```
QuantizationTable:
  Precision: 8-bit
  DestinationIdentifier: 1
Exif:
  ExifVersion: 0220
  FlashpixVersion: 0100
  ColorSpace: sRGB
  ComponentsConfiguration: 1, 2, 3, 0
  CompressedBitsPerPixel: 0.77
  PixelXDimension: 360
  PixelYDimension: 480
  DateTimeOriginal: 2006:06:02 10:00:19
  DateTimeDigitized: 2006:06:02 10:00:19
  ExposureTime: 0.001
  FNumber: 3.4
  ExposureProgram: unidentified
  ExposureBiasValue: 0
  MaxApertureValue: 2.6
  MeteringMode: pattern
  LightSource: unknown
  Flash: did not fire, auto mode
  FocalLength: 12.6
  FocalPlaneResolutionUnit: inches
  FileSource: DSC
  SceneType: directly photographed image
  CustomRendered: normal
  FocalLengthIn35mmFilm: 0
  SceneCaptureType: standard
  Saturation: normal
  Sharpness: normal
  SubjectDistanceRange: unknown
ApplicationSegments: APP0, APP1
```

JhoveView (Rel. 1.0, 2005-05-26)  
Date: 2006-08-09 16:03:51 BST  
**RepresentationInformation: C:\DROID\londontoexeter.htm**  
ReportingModule: HTML-hul, Rel. 1.1 (2005-04-22)  
LastModified: 2006-06-13 14:18:39 BST  
Size: 14883  
Format: HTML  
Status: Not well-formed  
ErrorMessage: TokenMgrError: Lexical error at line 1, column 22.  
Encountered: "\" (34), after : "  
MIMEtype: text/html

JhoveView (Rel. 1.0, 2005-05-26)  
Date: 2006-08-09 16:14:36 BST  
**RepresentationInformation: C:\DROID\excel.xls**  
ReportingModule: BYTESTREAM, Rel. 1.2 (2005-03-09)  
LastModified: 2004-08-04 05:00:00 BST  
Size: 5632  
Format: bytestream  
Status: Well-Formed and valid  
MIMEtype: application/octet-stream

Date: 2006-08-09 16:16:36 BST  
**RepresentationInformation: C:\DROID\project\_outputs.tbi**  
ReportingModule: ASCII-hul, Rel. 1.1 (2005-01-11)  
LastModified: 2006-01-23 16:57:20 GMT  
Size: 2236  
Format: ASCII  
Status: Well-Formed and valid  
MIMEtype: text/plain; charset=US-ASCII  
ASCIIMetadata:  
LineEndings: CRLF

JhoveView (Rel. 1.0, 2005-05-26)  
Date: 2006-08-09 16:19:14 BST  
**RepresentationInformation: C:\DROID\6to4svc.dll**  
ReportingModule: BYTESTREAM, Rel. 1.2 (2005-03-09)  
LastModified: 2004-08-04 05:00:00 BST  
Size: 100352  
Format: bytestream  
Status: Well-Formed and valid  
MIMEtype: application/octet-stream

JhoveView (Rel. 1.0, 2005-05-26)  
Date: 2006-08-09 16:22:56 BST  
**RepresentationInformation: C:\DROID\adojavas.inc**  
ReportingModule: ASCII-hul, Rel. 1.1 (2005-01-11)  
LastModified: 2004-08-04 05:00:00 BST  
Size: 14610  
Format: ASCII  
Status: Well-Formed and valid  
MIMEtype: text/plain; charset=US-ASCII  
ASCIIMetadata:  
LineEndings: CRLF

```
JhoveView (Rel. 1.0, 2005-05-26)
Date: 2006-08-09 15:35:14 BST
RepresentationInformation: C:\DROID\sndrec.wav
ReportingModule: WAVE-hul, Rel. 1.1 (2005-05-13)
LastModified: 2004-08-04 05:00:00 BST
Size: 58
Format: WAVE
Status: Not well-formed
SignatureMatches:
  WAVE-hul
ErrorMessage: Unexpected end of file
  Offset: 58
MIMEtype: audio/x-wave
```

```
JhoveView (Rel. 1.0, 2005-05-26)
Date: 2006-08-09 15:22:40 BST
RepresentationInformation: C:\DROID\unicodetxt.txt
ReportingModule: UTF8-hul, Rel. 1.1 (2005-01-11)
LastModified: 2006-08-01 14:05:41 BST
Size: 24
Format: UTF-8
Status: Not well-formed
ErrorMessage: UTF-16 little-endian encoding, not UTF-8
MIMEtype: text/plain; charset=UTF-8
```

