



CODA

**Curation of Digital Assets
2007**

Version 2:

English translation of parts from the final report (2009)

This work is subject to copyright. Copyright owner is LDB-centrum (English name LDP Centre). Permission is requested for every use of the material, fully or partially.

The reference LDB-centrum 2009 must always be included.

The work is not allowed to be reproduced, stored or transmitted in any form or by any means for commercial use.

© LDB-centrum 2009

PREFACE	- 4 -
1. INTRODUCTION	- 5 -
2. TEST OF PROGRAMS	- 6 -
2.1 DROID	- 6 -
2.2 JHOVE	- 7 -
2.3 TriID	- 8 -
2.4 FILE IDENTIFIER	- 8 -
2.5 CONCLUSION	- 9 -
APPENDIX - TEST OF PROGRAMS	- 10 -
SUMMARY OF RESULTS	- 11 -
TEST PROTOCOL DROID.....	- 12 -
TEST PROTOCOL DROID.....	- 13 -
TEST PROTOCOL JHOVE.....	- 14 -
TEST PROTOCOL JHOVE.....	- 15 -
TEST PROTOCOL TriID	- 16 -
TEST PROTOCOL TriID	- 17 -
TEST PROTOCOL FILE IDENTIFIER.....	- 18 -
TEST PROTOCOL FILE IDENTIFIER.....	- 19 -

Preface

The LDP Centre (Swedish name LDB-centrum) consists today of four parts working together with long-term digital preservation and access. The parts are:

National Archives of Sweden
National Library of Sweden
Lulea University of Technology
Municipality of Boden

The former fifth part, the Swedish National Archives of Recorded Sound and Moving Image, became in January 1st, 2009 a department of the National Library, “The Department of Audiovisual Material”.

The parts of the LDP Centre have co-operated since 2007 in a programme called CODA (Curation of Digital Assets). Every year a document will be published, describing the work done during the year. The document will be written in Swedish with an English abstract.

This document is an English version covering one chapter and an appendix from the report of 2007, translated to English in March 2009 by Allan Arvidson, National Library of Sweden.

Thank you Allan for your effort!

The document is free for non-commercial use. Even so, if you want to use the document, whether whole or part of the material, we appreciate if you inform us in advance. Always include a reference to us and our web page <http://www.ldb-centrum.se>

LDP Centre
Boden
Sweden

2009-03-24

1. Introduction

In 2007 the LDP Centre published the report “CODA, curation of digital assets”.

Quoting from their web site:

“The LDP Centre (Centre for Long-term Digital Preservation) is a competence centre for research and technical development and testing of methods and technologies for long-term digital preservation and access.”

<http://www.ldb-centrum.se>

The report contain, among other things, a test of some programs designed to identify file formats. It is witten in Swedish. In order to facilitate for the non-Swedish speaking reader a rough translation of the relevant parts has been made.

IT is NOT to be considered as the English translation of the original report, but merely as a help for the readers not mastering the Swedish language.

The original report, which can be found on the web site, is the authoritative document. Only chapter 7 and appendix 3 has been translated. Graphs may be missing, references ARE missing. Other omissions and simplification may occur. Any errors etc are solely due the translator.

Allan Arvidson
National Library of Sweden

2. Test of programs

(Chapter 7 in the original report)

A number of programs have been tested as to their ability to identify the format of a file. It was demanded that the programs should be able to identify the format and the version thereof. The programs tested were all downloaded from the web.

The following programs have been evaluated:

- DROID
- JHOVE
- TriID
- FILE IDENTIFIER

The tests have been made on a pc 2.8 Ghz 1GB ram and with Windows as operating system.

In order to get a good test material the logical formats used were examined with HHD Free Hex Editor as well as with the editor XVI32. This was done using the format registries PRONOM and WARC by extracting the formats internal signature. One of the above mentioned editors was then used to verify the internal hexadecimal signature of the test file.

The number of test files used was 16; eleven with text information, six image and four with moving images. The test was done twice on each file, one with and one without the file extension. This in order to see if the program only used the file extension or if it examined the internal structure of the file, or both.

2.1 DROID

DROID (Digital Record Object Identification) version 2.0 was developed by the National Archives in UK. It is a platform independent program written in Java with full documentation and a public API for easy integration with other systems. The program was built to identify file formats and this is done using the external and internal signature that the file has (magic number, file extension). These signatures are stored in an XML-structure which comes from PRONOM and is updated regularly. DROID has two user interfaces, a graphical Java swing GUI and a command line interface.

Input/output

Input is via a catalogue, files in the file system or via a simple XML-list. The result can be shown on screen, sent to a printer or generated as an XML-file or as a comma separated list.

The output consists of:

- PUID: Persistent Unique Identifier, PRONOM:s unique identifier for a file
- MIME: The MIME type
- Format: the name of the format
- Version: the version
- Status: status can be: positive (specific), positive (generic), tentative or negative

Positive (specific) means that the internal structure and the file extension matches a unique file format. If the internal structure matches a unique format but the extension doesn't a warning is given.

Positive (generic) means that the internal structure and the file extension matches a file generic file format. This can give several hits (several versions) if there is no unique internal structure for each version, only for the generic format. If the internal structure matches a generic format but the extension doesn't a warning is given.

Tentative is used when only the extension matches but with no hit on internal structure.

Negative means that the format couldn't be identified.

Warning: shows possible warnings, such as "possible file extension mismatch".

Unidentified: if the format couldn't be identified.

Result

In total DROID managed to identify 18 of 25 files correctly and 7 files partly correct. If the files didn't have an extension the result was lower, 15 of 25 files correctly, 5 partly correct. There were 5 files that DROID wasn't able to identify. One can here see that there were 5 files which DROID identified only by the extension. Without the extension it failed. If any of these files had had a wrong extension the result would have been worse, since DROID had misidentified the file.

DROID also gave more than one answer on three occasions (when testing without file extension). Summing up DROID gave the correct answer in 72% of the cases when testing with file extension and 60% without.

2.2 JHOVE

JHOVE-JSTOR/HARVARD Object Validation Environment version 1.1h was produced by Harvard University Library for the identification and validation of logical file formats. The program is written in Java, is platform independent and comes with full documentation. Every file format to be identified/validated has its own module written in Java. It is possible to write your own modules. At the time of writing there are 12 different modules. The communication with the program is via a graphical user interface (jhoveview.jar) or a command tool (jhoveapp.jar).

Input/output

The file(s) to be identified/validated can be feed to the program by catalog, in which all the files in the catalogue will be treated. It's also possible to process a single file. The result is shown on the screen, as an XML structured file or as a text file. An audit file can also be generated. The information generated is very comprehensive: file name, version, status as well as a lot of metadata about the file. In this report we've concentrated on the file name and version.

Result

The test was done both using the graphical user interface and the command line interface. In total JHOVE identified 8 of the 25 files correctly. No difference was found between testing with and without the file extension. This shows that JHOVE only uses the internal structure of a file for its analysis. As the program is built with 12 standard modules, one for each format, the result can almost be predicted. When JHOVE could identify the format it did a very good job. Apart from pure identification also the validation and extraction of metadata. It identified 32% of the files correctly.

2.3 TriID

TriID-File identifier version 2.02 is a tool to identify files from their bit stream. The program was developed by Marco Portello and works on Windows 32 bit and Linux x86. The identification is done by comparing the internal file pattern against a given pattern. It's possible to train the program by running it against a selection of files and update the database. Communication with the program is via a command window (trd.exe) or via a graphical windows application (TriDNet.exe).

Input/output

Files can be given to the program as a catalog or as individual files. The result is only shown on screen and it's possible to get several hits. If there are more than one hit the result is a list with probabilities in percent. The output is simple, only format, no version.

Result

In total TriID identified 7 files correctly and 13 partly correct. The problem is that the program doesn't handle versions of a format which is one of the criteria in the test. Also giving the answer with a percent probability can be problematic since it leaves the answer open to interpretation. No difference was noted when test with and without the file extension, confirming that identification only uses the internal structure of a file. TriID identified 28% of the 25 files correctly.

2.4 File identifier

File identifier 0.6.1 (beta version) was developed to identify files via the internal structure and extract a small amount of metadata. The program was developed by Optima SC Inc. and the version tested here is freeware. Platforms supported are Windows 32 bit and Linux x86. Communication is via the command window (file.exe) and at the time of writing it supports some 600 formats.

Input/output

Files can be processed by catalogue, ie all the files in a catalogue, or file by file. The output is:

- File name: The name of the format and version
- File class: Text, image, audio or moving pictures
- MIME: The MIME type
- File path: Absolute file path
- Metadata: Some metadata such as creation date, modification date and some other metadata depending on file class

The result is shown on screen, as an HTML document or an SFV report.

Result

File identifier managed to identify 8 files with correct format and version. 12 files were partly identified, either without version or only a generic answer was given. There was no difference between testing with or without the file extension. The score in identifying formats was 32%.

2.5 Conclusion

This test was performed in order to determine if a program could identify a file format and version. The best were DROID and JHOVE. The problem with the other programs were that they couldn't deliver the version or that the result wasn't unambiguous.

DROID scored best with 72% correctly identified files with the file extension and 60% without extensions, showing that even the best program isn't perfect.

Looking at the lower result, without file extensions, where the identification is done only using the internal structure which is what one should do if one is not sure that the extension is correct. The result for DROID is a little more the half of the files. JHOVE only handles 12 different formats, but these were perfectly identified, validated and metadata was extracted. It's also possible to write your own format modules. If you know roughly which formats you have and they coincide with JHOVE:s formats, this program is a good choice.

However, more tests are needed on a larger collection of files, how to handle the output from the programs, what information is important and how it is to be used. This should be known before using the tools in production.

Appendix - Test of programs

(Appendix 3 in the original report)

This appendix contains the material used, which formats have been used, diagrams and test protocols.

Files

These files were used in the identification test. The files were divided into the categories: text, image, audio and moving image. There is also a file with embedded audio. It was placed in the category “text”.

Text

- Plain text Europe iso 8859-1 (txt). Created by MS Word 2003
- Plain text unicode-8 (txt). Created by MS Word 2003
- Rich text format (rtf) v1.4. Created by Openoffice 2.0
- MS Word 2.0 (doc) v2.0. Created by MS Word 2.0
- MS Word 6.0 (doc) v6.0. Created by MS Word 6.0
- MS Word 2003 (doc) v8.0. Created by MS Word 2003
- MS Word 2003 with embedded audio file (doc) v8.0. Created by MS Word 2003
- MS Works 6.0/70 (wps). Created by MS Word 2003
- MS Excel spread sheet 4.0 (xls) v4.s. Created by MS Excel
- Portable document format (pdf) v.4 (pdf/A 1-b). Created by Adobe Acrobat professional
- Openoffice text file (odt) v1.0. Created by Openoffice 2.0

Image

- Tiff 2.2 (Exchangeable Image file format (uncompressed) 2.2) (tif) v2.2. Created by digital camera
- Gif 89a (gif) v.89v. File created by converting in Photoshop CS2 from a tiff 2.2 image
- JPEG/JIFF (jpg) v1.02. File created by converting in Photoshop CS2 from a tiff 2.2 image
- JPEG 200 part 1. File created by converting in Photoshop CS2 from a tiff 2.2 image
- Windows OS/2 bitmap graphics (bmp) v3.0. File created by converting in Photoshop CS2 from a tiff 2.2 image
- Portable network graphics (png) v1.1

Audio

- mpeg-1 layer 3 (mp3). Downloaded from web
- Wave form audio (wav). File from operating system
- Standard MIDI file format (mid) (no version). File from operating system
- Windows media audio file (mwa). File from operating system

Moving image

- Quicktime (qt, mov) v3.0. Created by digital camera
- mpeg-1 video (mpg, mpeg)(no version) File created by converting from QT to mpeg-1
- mpeg-2 video (mpg) (no version). File created by converting from QT to mpeg-1
- Audio Video interleave file (avi). Downloaded from the web

Summary of results

This is a summary of the identification tests. 25 files were used in the test. The criteria used were:

- Yes, correct format and version
- Partly, correct format
- No, wrong on both format and version

Two numbers are given in each cell in the table below. The one within parenthesis is when the test was made without the file extension. The number in parenthesis shows the result when the test was made with file extension. Both absolute numbers and percent are shown.

	Correct		Partly correct		Wrong		Sum
	Number	%	Number	%	Number	%	
DROID	18 (15)	72 (60)	7 (5)	28 (20)	0 (5)	0 (20)	25
JHOVE	8 (8)	32 (32)	0 (0)	0 (0)	17 (17)	68 (68)	25
TrilD	7 (7)	28 (28)	13 (13)	52 (52)	5 (5)	20 (20)	25
File Id	8 (8)	32 (32)	12 (12)	48 (48)	5 (5)	20 (20)	25

Test protocol DROID

Tool: DROID

Test: identify/extract metadata

With file extension

File format	Yes	Partly	No
Txt 8859-1		X(1)	
Txt unicode-8		X(1)	
Rtf v1.4		X(7)	
Doc v2.0	X		
Doc v6.0	X		
Doc v8	X(5)		
Doc v8 + audio file		X(4)	
wks	X(5)		
Xls v4.0s	X		
Pdf v1.4	X		
Odt v1.0	X		
Tif v2.2	X		
Gif v89a	X		
Jpg v1.02	X		
Jp2 part1	X(6)		
Bmp v3.0	X		
Png v1.1	X		
mp3	X(6)		
wav	X		
mid	X(6)		
wma	X		
Qt v3.0		X(2)	
mpeg-1		X(3)	
mpeg-2		X(3)	
avi	X		
Total	18	7	0

Comments:

1. Returns several answers (tentative)
2. No version, correct format
3. Several answers mpeg1/mpeg2
4. Identifies the format, but not the embedded audio file
5. Returns OLE2 compound document
6. Correctly identified, but gives warning (tentative)
7. Returns several answers (generic)

Test protocol DROID

Tool: DROID

Test: identify/extract metadata

Without file extension

File format	Yes	Partly	No
Txt 8859-1			X
Txt unicode-8			X
Rtf v1.4		X(5)	
Doc v2.0	X(1)		
Doc v6.0	X(1)		
Doc v8	X(1)		
Doc v8 + audio file		X(6)	
wks	X(1)		
Xls v4.0s	X(1)		
Pdf v1.4	X(1)		
Odt v1.0	X(1)		
Tif v2.2	X(1)		
Gif v89a	X(1)		
Jpg v1.02	X(1)		
Jp2 part1			X(2)
Bmp v3.0	X(1)		
Png v1.1	X(1)		
mp3			X(2)
wav	X(1)		
mid			X(2)
wma v2.0	X		
Qt v3.0		X(3)	
mpeg-1		X(4)	
mpeg-2		X(4)	
avi	X(1)		
Total	15	5	5

Comments:

1. Warning on file extension
2. Couldn't identify format, managed in previous test with file extension
3. Couldn't identify version
4. Several answers mpeg1/mpeg2
5. Returns several answers (generic)
6. Identifies the format, but not the embedded audio file

Test protocol JHOVE

Tool: JHOVE

Test: identify/extract metadata

With file extension

File format	Yes	Partly	No
Txt 8859-1	X(5)		
Txt unicode-8	X		
Rtf v1.4			X(4)
Doc v2.0			X(1)
Doc v6.0			X(1)
Doc v8			X(1)
Doc v8 + audio file			X(1)
wks			X(1)
Xls v4.0s			X(1)
Pdf v1.4	X		
Odt v1.0			X(1)
Tif v2.2	X(2)		
Gif v89a	X		
Jpg v1.02	X		
Jp2 part1	X		
Bmp v3.0			X(1)
Png v1.1			X(1)
mp3			X(1)
wav	X(3)		
mid			X(1)
wma			X(1)
Qt v3.0			X(1)
mpeg-1			X(1)
mpeg-2			X(1)
avi			X(1)
Total	8		17

Comments:

1. Returns "well-formad and valid", says it's a byte stream
2. Tiff 6.0 file, signature vs t2.2, profile exif 2.2
3. Says it's schema version 1.02b
4. Says it's an ascii file
5. Says it's us-ascii

Test protocol JHOVE

Tool: JHOVE

Test: identify/extract metadata

Without file extension

File format	Yes	Partly	No
Txt 8859-1	X(5)		
Txt unicode-8	X		
Rtf v1.4			X(1)
Doc v2.0			X(4)
Doc v6.0			X(1)
Doc v8			X(1)
Doc v8 + audio file			X(1)
wks			
Xls v4.0s			X(1)
Pdf v1.4	X		
Odt v1.0			X(1)
Tif v2.2	X(2)		
Gif v89a	X		
Jpg v1.02	X		
Jp2 part1	X		
Bmp v3.0			X(1)
Png v1.1			X(1)
mp3			X(1)
wav	X(3)		
mid			X(1)
wma			X(1)
Qt v3.0			X(1)
mpeg-1			X(1)
mpeg-2			X(1)
avi			X(1)
Total	8		17

Comments:

1. Returns "well-formed and valid", says it's a byte stream
2. Tiff 6.0 file, signature t6 vs t2.2, profile exif 2.2
3. Says it's schema version 1.02b
4. Says it's an ascii file
5. Says it's us-ascii

Test protocol TrilD

Tool: TrilD

Test: identify/extract metadata

With file extension

File format	Yes	Partly	No
Txt 8859-1			X
Txt unicode-8			X
Rtf v1.4		X(1)	
Doc v2.0			X
Doc v6.0		X(2)	
Doc v8		X(2)	
Doc v8 + audio file		X(2)	
wks		X(3)	
Xls v4.0s			X
Pdf v1.4		X(1)	
Odt v1.0		X(4)	
Tif v2.2		X(5)	
Gif v89a	X(6)		
Jpg v1.02		X(7)	
Jp2 part1	X(8)		
Bmp v3.0			X(9)
Png v1.1		X(1)	
mp3	X		
wav	X(10)		
mid	X		
wma	X(11)		
Qt v3.0		X(1)	
mpeg-1		X(12)	
mpeg-2		X(13)	
avi	X(14)		
Total	7	13	5

Comments:

1. Gives correct name, but no version
2. Says 80% MS Word doc.; alt: generic OLE2/multi stream compound file
3. 84% work file, no version; alt generic OLE2/multi stream compound file
4. 80% opendocument text, alt OD spreadsheet
5. Says tiff (little endian), no version
6. Correct answer 60%, 30% gen GIF, 30% ??
7. Jfif-exif jpeg bitmap 33% no version, jfif jpeg bitmap 25%, jpeg bitmap 20% etc
8. Jpeg-2000 bitmap 91%
9. Run length encoded bitmap (rle) 51%, win bmp 49%
10. riff/wave std-audio 50%, gen riff container
11. WMAudio 50%, WMVideo 50%
12. mpeg video, doesn't distinguish between 1 and 2
13. mpeg video 75%, BONK lossy/lossless
14. Avi 51%, 49 gen RIFF container

Test protocol TrilD

Tool: TrilD

Test: identify/extract metadata

Without file extension

File format	Yes	Partly	No
Txt 8859-1			X
Txt unicode-8			X
Rtf v1.4		X(1)	
Doc v2.0			X
Doc v6.0		X(2)	
Doc v8		X(2)	
Doc v8 + audio file		X(2)	
wks		X(3)	
Xls v4.0s			X
Pdf v1.4		X(1)	
Odt v1.0		X(4)	
Tif v2.2		X(5)	
Gif v89a	X(6)		
Jpg v1.02		X(7)	
Jp2 part1	X(8)		
Bmp v3.0			X(9)
Png v1.1		X(1)	
mp3	X		
wav	X(10)		
mid	X		
wma	X(11)		
Qt v3.0		X(1)	
mpeg-1		X(12)	
mpeg-2		X(13)	
avi	X(14)		
Total	7	13	5

Comments:

1. Gives correct name, but no version
2. Says 80% MS Word doc.; alt: generic OLE2/multi stream compound file
3. 84% Work file, no version; alt generic OLE2/multi stream compound file
4. 80% opendocument text, alt OD spreadsheet
5. Says tiff (little endian), no version
6. Correct answer 60%, 30% gen GIF, 30% ??
7. Jfif-exif jpeg bitmap 33% no version, jfif jpeg bitmap 25%, jpeg bitmap 20% etc
8. Jpeg-2000 bitmap 91%
9. Run length encoded bitmap (rle) 51%, win bmp 49%
10. Riff/wave std-audio 50%, gen riff container
11. WMAudio 50%, WMVideo 50%
12. mpeg video, doesn't distinguish between 1 and 2
13. mpeg video 75%, BONK lossy/lossless audio comp.
14. Avi 51%, 49 gen RIFF container

Test protocol File identifier

Tool: file identifier

Test: identify/extract metadata

With file extension

File format	Yes	Partly	No
Txt 8859-1			X
Txt unicode-8			X
Rtf v1.4		X(1)	
Doc v2.0			X
Doc v6.0		X(2)	
Doc v8		X(2)	
Doc v8 + audio file		X(3)	
wks		X(2)	
Xls v4.0s			X
Pdf v1.4		X(1)	
Odt v1.0			X(4)
Tif v2.2		X(1)	
Gif v89a	X		
Jpg v1.02	X		
Jp2 part1	X		
Bmp v3.0		X(1)	
Png v1.1		X(1)	
mp3	X		
wav	X		
mid	X		
wma	X		
Qt v3.0		X(1)	
mpeg-1		X(5)	
mpeg-2		X(5)	
avi	X		
Total	8	12	5

Comments:

1. Name of format and MIME type found, but not the version
2. Designated MS OLE compound document
3. Designated MS OLE compound document, embedded audio file not found
4. Says the file is a pzip archive file
5. Returns mpeg multimedia (image/audio) streaming file, do not distinguish between mpeg-1 and mpeg-2

Test protocol File identifier

Tool: file identifier

Test: identify/extract metadata

Without file extension

File format	Yes	Partly	No
Txt 8859-1			X
Txt unicode-8			X
Rtf v1.4		X(1)	
Doc v2.0			X
Doc v6.0		X(2)	
Doc v8		X(2)	
Doc v8 + audio file		X(3)	
wks		X(2)	
Xls v4.0s			X
Pdf v1.4		X(1)	
Odt v1.0			X(4)
Tif v2.2		X(1)	
Gif v89a	X		
Jpg v1.02	X		
Jp2 part1	X		
Bmp v3.0		X(1)	
Png v1.1		X(1)	
mp3	X		
wav	X		
mid	X		
wma	X		
Qt v3.0		X(1)	
mpeg-1		X(5)	
mpeg-2		X(5)	
avi	X		
Total	8	20	5

Comments:

1. Name of format and MIME type found, but not the version
2. Designated MS OLE compound document
3. Designated MS OLE compound document, embedded audio file not found
4. Says the file is a pzip archive file
5. Returns mpeg multimedia (image/audio) streaming file, do not distinguish between mpeg-1 and mpeg-2



LDP Centre
Centre for Long-term Digital Preservation

Skapa Företagsby
Teknikvägen 3-13
961 50 BODEN
Sweden

Phone: +46 921 573 00
E-mail: kontakt@ldb-centrum.se
Web: <http://www.ldb-centrum.se>