



CODA-META

**Curation of Digital Assets - Metadata
2008**

***Version 2:
English translation (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

Abstract

This report describes the evaluation of four tools for metadata extraction. The tools have been evaluated by testing them on a total of 27 files of different formats and versions. The analyze was made using the following norms, produced by the CODA group:

- General technical and describing metadata
- Technical metadata for video formats
- Technical metadata for sound files
- Technical metadata for the format .tiff
- Technical metadata for text formats

Test results were compared and the metadata tools were ranked. The best software turned out to be Exiftool which extracts metadata from a majority of the files tested against predetermined protocols.

ABSTRACT.....	- 3 -
PREFACE.....	- 5 -
INTRODUCTION.....	- 6 -
1. PROJECT CODA-META	- 7 -
1.1 AIM	- 7 -
1.2 TARGET	- 7 -
1.3 DISPOSITION.....	- 7 -
1.4 WORKING TEAM	- 7 -
2. THE PROBLEM.....	- 8 -
3. TEST OF TOOLS	- 8 -
3.1 METHOD.....	- 8 -
3.2 LIMITATIONS	- 9 -
3.3 EXIFTOOL	- 9 -
3.4 JHOVE.....	- 10 -
3.5 METADATA EXTRACTION TOOL	- 10 -
3.6 FILE IDENTIFIER	- 11 -
4. LIST OF GENERAL METADATA FROM KB	- 12 -
4.1 ANALYSIS OF THE DISTRIBUTION CATEGORY TOTAL.....	- 12 -
4.2 ANALYSIS OF TECHNICAL AND DESCRIPTIVE METADATA.....	- 12 -
4.3 ANALYSIS OF TECHNICAL AND DESCRIPTIVE METADATA ON A FEW FORMATS	- 13 -
4.4 TESTING ABILITY TO OPEN AND READ FILES	- 14 -
4.5 SUMMARY, TEST AGAINST THE KB LIST OF GENERAL METADATA	- 14 -
5. TEST ON SPECIFIC FORMATS.....	- 15 -
5.1 TEST OF THE TIFF FORMAT.....	- 15 -
5.2 TEST OF TEXT FORMATS	- 16 -
5.3 TEST ON VIDEO AND AUDIO FORMATS	- 17 -
5.4 CONCLUSION OF THE TESTS ON SPECIFIC FORMATS	- 18 -
6 CONCLUSIONS	- 19 -
APPENDICES	- 20 -
APPENDIX 1: FILE LISTS AND TEST FILES	- 20 -
APPENDIX 2: KB LIST OF GENERAL METADATA.....	- 22 -
APPENDIX 3.....	- 23 -
APPENDIX 4: TECHNICAL METADATA FOR AUDIO.....	- 24 -
APPENDIX 5: TECHNICAL METADATA FOR TIFF	- 25 -

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 from the report of 2008, 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

2009-04-02

Introduction

In 2008 the LDP Centre published the report “CODA-META, Curation of Digital Assets - Metadata”.

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 contains a test of some programs designed to extract metadata from different file formats. It is written 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. Some things have been excluded: graphs, references. Other omissions and simplification may occur. Any errors etc are solely due the translator.

Allan Arvidson
National Library of Sweden

1. Project CODA-META

The CODA project is a collaboration between the LDP Centre (Centre for Long-term Digital Preservation), the National Archives of Sweden (RA), the National Library of Sweden (KB) and the Swedish National Archive of Recorded Sound and Moving Images (SLBA). Planning for the project was made in the second half of 2006 and the project started around the turn of 2006/2007. There were two different projects, CODA-POP and CODA-FORM, both of which involved all partners.

Test and evaluation of tools for extraction of metadata from files was an item which was not finished in time for the FORM report. It was decided during the partners meeting in January 2008 to finish this work during 2008. This project, named CODA-META, should finalize this point.

1.1 Aim

The aim of CODA-META is to evaluate and rank four different tools for extracting metadata.

1.2 Target

The report is primarily aimed to technical and administrative staff at KB, SLBA and RA. Since large parts of the report is written on a generic level, some of the text can also be of interest for personnel in other organisations working with long term digital preservation.

1.3 Disposition

Chapter 2 describes the value of being able to extract metadata from files. Chapter 3 describes, in general terms, methods, tools and limitations of the test and gives information about the different tools. In chapter 4 the tools are tested against the list of general metadata from KB. The results is analysed and presented in tables and in a summary for the chapter. In chapter 5 the tools are tested against lists of technical metadata for a smaller number of formats. The result is examined and analysed and presented in tables. In chapter 6 there is a summary and final analysis for the whole list of metadata, the tools are ranked and their pros and cons are described.

1.4 Working team

In the work with CODA-META the following persons participated:

Stina Degerstedt, KB, Project Leader

Göran Lindqvist, LDB, Project Administrator

Lena Lindbäck, LDB

Jan Aspenfjäll, LDB

Allan Arvidson, KB

Martin Jacobson, SLBA

Göran Konstenius, SLBA

Magnus Geber, RA

Johan Ekdahl, RA

The work was carried out from February to September 2008. Five Marratech meetings were held as well as a workshop in Stockholm. All testing was done by LDB Centre in Boden.

2. The problem

To ensure that a file is accessible in the long term it is necessary to know its technical organisation, content, dependencies on other digital material and to know what has happened to the file over time. It is important to have metadata, e.g. information about the object.

Metadata can be divided into: technical, descriptive and administrative metadata. Some of this information is stored inside the file. The amount of metadata stored in a file differs a lot between different formats. One way to get metadata from a file is to use a tool which extracts metadata from the file.

3. Test of tools

The tests are performed in order to evaluate the possibility to extract metadata from files. The demands are that the tools should extract the metadata listed in KB:s list. The tools ability extract specific technical metadata is tested for a smaller number of formats. The tools evaluated are all open source tools freely available:

- Exiftool
- Jhove
- Metadata extraction tool
- File identifier

The computer used for the tests was 2.8 Ghz pc with 1 GB ram running Windows XP and with MS Office installed. This computer was used for all tests.

A total of 27 files were tested. Of these 16 were categorised as follows:

- Text (7)
- Image (7)
- Sound/Video (2)

The remaining 11 files were used in test of specific technical metadata. The choice of file types is based on the format lists produced for the CODA-FORM project by the participating institutions. The files were produced in a number of different ways, by scanner, digital cameras, programs and by converting files from other formats.

In 16 of the files the metadata was enriched to get more metadata in the test files. The files were then examined using a hexadecimal editor (HxD hexeditor) to get metadata that was not extracted by the programs. All the metadata in a file was then documented to have it available when testing.

3.1 Method

The tools are ranked against a number of lists of metadata from the CODA group. The lists comprises the following:

- General technical and descriptive metadata from KB
- Technical metadata for TIFF
- Technical metadata for text formats
- Technical metadata for sound formats
- Technical metadata for video formats

First an assessment is made against the list of general metadata from KB which will examine general and descriptive metadata. Then an extend check is made of which technical metadata the tools deliver. The different formats for text, video and audio as well as TIFF is checked against the specific list of metadata for the format in question. The result is then compiled. This compilation is the basis for the final assessment of the tools.

3.2 Limitations

It's important to know the limitations, first concerning the test of the list of general metadata from KB. This gives the tools ability to satisfy a given list of metadata with a limited number of files. The test of specific technical metadata examines a couple of different formats and versions thereof of formats such as tiff, mpeg and text files. Here there is also a limitation coming from the technical test protocols. All results are given with respect to these test protocols.

3.3 Exiftool

The program Exiftool is a platform independent perl library and a command tool. There is also a Windows executable for the Microsoft platform as well as a Macintosh OS X package. The program can read, write and edit metadata information in image, sound and video files. It supports a wide range of different kinds of metadata such as:

EXIF	GPS	IPTC	XMP
JFIF	GeoTIFF	ICC profil	Photoshop IRB
Flashpix	AFCP	ID3	

The program can also read metadata from cameras:

Canon	Casio	Fujifilm	Hewlett Packard
JVC/Victor	Kodak	Leaf	Minolta/Konica
Nikon	Olympus/Epson	Panasonic/Leica	Pentax/Asahi
Ricoh	Sanyo	Sigms/Foveon	Sony

The program was developed by Phil Harvey, the licens is GNU. It's possible to modify the perl script for specific applications. There is a certain amount of development going on.

Input/output

The program takes its input from the command line, from a separate file or even a whole catalogue. The output from the program can be as a text file, a HTML file or on screen. The information generated can also be sorted and structured in various ways, e. g. after group (EXIF, XMP etc). It is also possible to limit the information to a certain category (e. g. only IPTC) via parameters to the command tool. The output is delivered as a colon separated list, alternatively a tab separated list. More about this can be found in the program help. The version of Exiftool tested here is 7.30.

3.4 Jhove

Jhove-jstor/Harvard Object Validation Environment version 1.1 is a tool developed by the Harvard University Library for identification, validation and extraction of metadata from files. It is written in Java and is platform independent and comes with full documentation.

Every file format has its own module, written in Java, which identifies, validates and extracts metadata. At the time of writing there are 12 modules for different formats.

In the table below 11 modules are presented:

Text	Image	Audio/Video	Markup language
Ascii	Aiff	Wave	HTML
Utf-8	Tiff		XML
PDF	Gif		
	Jpeg		
	Jpeg2000		

If Jhove fails to identify the format it chooses the 12:th module, the bytestream module. As the program and the source code is available under GNU Lesser General Public License (LGPL) it's possible to develop your own modules in Java.

The program is actively developed and a new version is being developed. Communication with the program is via a graphical interface (jhoveview.jar) or a command line interface (jhoveapp.jar).

Input/output

The file or files to be processed can be given to Jhove as a catalogue, all the files in a catalogue or as individual files. The result is given on the screen or as a HTML file or as a regular text file. An audit file can also be generated. The information generated is very detailed: file name, version, status and lots of metadata about the file.

3.5 Metadata extraction tool

Metadata extraction tool was developed by Sytec Resources for the National Library of New Zealand in 2003. In 2007 it was decided to release the software as open source. The program is written to extract metadata from files for digital presentation.

It's written in Java and has a command line interface for Unix and a graphical interface for Windows as well as a command line interface.

The program comes with good documentation, installation guide, information about the structure of the software and a user guide. The source code is available so it's possible to develop your own modules. The program has a number of format modules which extracts metadata from files.

The file formats supported by Metadata extraction tool are:

Text documents	Image	Audio/Video	Markup languages
MS Word v 2-6	Tiff	Wav	HTML
MS Power point	Jpeg	Mp3	XML
MS Excel	Bmp		
MS Works	Gif		
Word perfect			
Open Office v 1.0			
PDF			

Input/output

Files can be feed via the graphical interface or by giving the program a list of files. Metadata can be extracted from a single file, a catalogue or all the file in a list. The resulted is presented on screen or as a XML-file. The data can be presented in two ways: each file in its own XML-file (one for each file) or the results from all processed files in a single output file.

As the source code is available it's expected that some development is going on. The latest version is 3.1GA and that is the one tested here.

3.6 File identifier

File identifier 0.6 (beta version) is was made to identify and extract metadata from file formats. It was developed by Optima SC Inc. The version tested here is the freeware version. File identifier runs on Windows 32 bit machines and Linux x86. The program is run from the command line (file.exe). Today about 600 file formats are supported for identification. Metadata can be extracted from about 30 different formats.

Input/output

Files can be processed by catalogue, e. g. all files in a catalogue, or file by file. The output is: file name, class of file, mime type, path to file and some metadata about the file such as creation data, date of modification and some specific data depending on the class of format. The result is presented on screen, as an XML-file or as an SFV-report.

4. List of general metadata from KB

In this test the programs are tested for ability to extract metadata elements given in the KB list. The list is divided into general technical metadata and descriptive metadata. The two last tags in the list, “if the file has been truncated” and “checksum and algorithm” were not included in the test.

In total 16 files were used in the test. The files were divided into the categories: text/html, image and audio and video.

Four types of analysis has been done:

- The distribution per category /total
- Distribution of technical and descriptive metadata / total
- Distribution of technical and descriptive metadata / total for tiff, pdf/a and mp3
- Test of the programs ability to open/read the files

4.1 Analysis of the distribution category total

An analysis has been made in order to judge which category of files a certain tool handles well. Exiftool is best at extracting metadata from the KB list with a hit rate of 52%. Metadata extraction tool manages 34%, Jhove 32% and File identifier only 28%, see table below.

A calculation has been done for each category according to the formula:
(sum of all tags found in all protocols for a given category) / (all tags for a given category) *
(the number of protocols). This will give a value for each category.

Result in table below:

	Exiftool	File identifier	Jhove	MetaExtr tool
Text	46%	22%	34%	42%
Image	62%	35%	36%	29%
Audio/Video	41%	23%	9%	27%

According to this table it is easiest to get metadata for the category “image”, the result being 62% for Exiftool, Jhove 36%, file identifier 35% and Metadata extractor tool 29%. Next comes the category “text”. It's hardest to get metadata for “Audio/Video”. Here the best tool, Exiftool, manages 41% and the others even worse.

4.2 Analysis of technical and descriptive metadata

An analysis has been made of the distribution of general technical metadata and descriptive metadata on all 16 files.

Exiftool has an even distribution with 24% on technical metadata and 28 % on descriptive metadata. File identifier, Jhove and Metadata extractor show a greater difference between technical and descriptive metadata, se the table below:

	Exiftool	File identifier	Jhove	MetaExtTool
Technical	24%	10%	20%	27%
Descriptive	28%	18%	12%	7%

It is interesting to note that Metadata extractor tool is the best at extracting technical metadata from the KB list, but worse when it comes to descriptive metadata.

4.3 Analysis of technical and descriptive metadata on a few formats

In this analysis three formats has been used, one from each category. The following have been used:

- Text (PDF/A)
- Image (Tiff)
- Sound and video (mp3)

These formats have been chosen because nearly all programs can read them with the exception for mp3 and Jhove. It's interesting to check how much metadata they can extract from a limited number of files.

		Exiftool	File identifier	Jhove	MetaExtTool
TIFF	techn	27.3%	9.1%	36.4%	36.4%
	descr	45.5%	36.4%	45.5%	9.1%
PDF/A	techn	27.3%	9.1%	36.4%	36.4%
	descr	45.5%	27.3%	45.5%	9.1%
MP3	techn	27.3%	9.1%	9.1%	36.4%
	descr	27.3%	27.3%	0.0%	9.1%

The table shows that Jhove can extract 82% of all metadata for tiff and pdf/a which is a good result. However, it's not so good for mp3. Jhove can only handle a limited number of formats and if there is problem with the validation of a file, the resulting data is almost zero.

Exiftool manages 73% on tiff and pdf/a and 55 % for mp3. This is lower than Jhove. However, we know from previous examinations that Exiftool manages a wider range of formats.

Metadata Extractor is good on technical metadata, about the same as Jhove. From the previous investigations we know however that it can handle more formats than Jhove. File identifier shows the lowest score. Its best result is for tiff with only 46 % of the items on the KB list.

File identifier is good at reading different files, 14 out of 16 files. However it produces very little metadata as can be seen in all the tests above.

4.4 Testing ability to open and read files

Finally we present a table showing which ones of the 16 files the different programs could open and read. Exiftool and File identifier could open and read 14 files while Metadata extractor managed 13 files and Jhove only 7 files. Open office text file caused the greatest problems, only Metadata extractor could read this format. No file was unreadable for all programs.

Files	Exiftool	File identifier	Jhove	MetaExtTool
Text file ut8	n	n	y	y
Html L4.01	y	y	n	y
Html L1.0	y	y	n	y
MS word	y	y	n	y
Open office text	n	n	n	y
PDF/A-1b	y	y	y	y
PDF v1.3	y	y	y	y
Tiff 6.0	y	y	y	y
Tiff 6.0, EXIF 2.2	y	y	y	y
Jpeg v1.01	y	y	n	y
Jpeg v1.02	y	y	n	y
Jpeg 2000	y	y	y	n
Png v1.1	y	y	n	n
Gif 89a	y	y	y	y
MPEG layer 3	y	y	n	y
Mpeg 1 Video	y	y	n	n

4.5 Summary, test against the KB list of general metadata

In conclusion, Exiftool is best in handling the 16 files in this test. It only failed to open the pure text file (ut8) and Open office text document. It got the best results in the test described in chapter 4.1 with a score of 52% against the KB list and even distribution of technical and descriptive metadata.

Jhove Scores well in the test “4.3 Analysis of technical and descriptive metadata on a few formats”. When the program can open and read the file it gives a lot of metadata. The problems are: it handles a limited number of formats and if it fails to validate the file it gives very little, if any, metadata.

Metadata Extractor Tool is the best to extract general technical metadata in the KB list. This can be seen in sections 4.2 and 4.3. It managed to open 13 of the 16 files used in the tests. It was the only program that managed to open the Open Office text file. Descriptive metadata is however not its strongest point.

5. Test on specific formats

The tests described here are directed towards technical metadata. We're testing tiff and some simple text files. Some video and audio formats are also tested against the technical test protocol.

5.1 Test of the tiff format

Seven files have been used in this analysis. The programs are tested against a test protocol (app. 5) consisting of a general technical part and an extended technical metadata part.

The general part tests:

- Which format
- Which version
- Possible coding
- File size

The extended part checks the following tags in the tiff file.

Name of tag	Number
NewSubfileType	254
SubfileType	255
BitsPerSample	258
Compression	259
PhotometricInterpretation	262
Thresholding	263
SamplesPerpixel	277
Xresolution	282
PlanarConfiguration	284
IPTC/NAA	33723

As can be seen in the table below Metadata Extractor Tool has a score of 100% for all tags in the general part for all files. Jhove 70% and Exiftool 75%. File identifier has the lowest score with only 25%.

In the first 4 rows in the table below, which corresponds to the general technical metadata, it can be seen that it's the version of the format that causes problems for Exiftool. Reading all the files is Jhove's weak point, as can be seen by the fact that 71% of the files could only be read, except for the file size.

Moving on to the extended list of metadata given in the table above. Here Exiftool is best with a score of 70%, followed by Jhove (60%) and Metadata Extractor tool (40%). File identifier has the lowest score with 10%. Looking closer at the table it can be seen that SubfileType (255) isn't found by any program. That is because NewSubfileType (254) has been used instead. Comparing the two best programs in this test, Jhove and Exiftool; we can see from the table below that Jhove manages to extract information to all the tags.

However, it only managed to open 7 of the 9 files (71%). The result for the tag IPTC/NAA (33723) is 29%. However only 2 of the 9 files had any metadata corresponding to this tag, which is 29%.

Exiftool failed to extract any metadata for Thresholding (263) This program is however not so limited in file formats as Jhove and has a better score when it comes to the number of files it can handle. Metadata extractor tool gets a lower score when tested against the extended metadata. It only manages NewSubfileType (254), BitsPerSample (258), Compression (259) and Xresolution (282). Failing on the other 4 tags. File identifier does so badly that we don't consider it in this test.

Summing up the result is as shown in diagram 4: Exiftool 70%, Jhove 65%, Metadata extractor tool 57% and File identifier 14%.

Program	Exiftool	File identifier	Jhove	MetaExtTool
Which format	100%	100%	71%	100%
Which version	0%	0%	71%	100%
Coding (eg BASE64)	100%	0%	71%	100%
Size	100%	0%	100%	100%
NewSubfileType (254)	100%	0%	71%	100%
SubfileType (255)	0%	0%	0%	0%
BitsPerSample (258)	100%	0%	71%	100%
Compression (259)	100%	0%	71%	100%
PhotometricInterpretation (262)	100%	0%	71%	100%
Thresholding (263)	0%	0%	71%	0%
SamplesPerPixel (277)	100%	0%	71%	0%
Xresolution (282)	100%	100%	71%	100%
PlanarConfiguration (284)	71%	0%	71%	0%
IPTC/NAA (33723)	29%	0%	29%	0%

5.2 Test of text formats

In this test 5 simple text documents are used in a test against the test protocol. The following tags, for general technical metadata, are tested:

- Which format
- Which version
- Coding
- File size

For the extended test the following tags are selected:

- Character representation
- Coded numerical formats
- Fixed post/row size

Only two programs, Jhove and Metadata extractor tool, can read simple text files. Jhove also manages to give extended technical metadata as can be seen in table 10. However the program only manages to open 3 of the 5 files and only found the tags corresponding to character representation in the extended part. This is better than the other programs, but still rather low.

Program	Exiftool	File identifier	Jhove	MetaExtTool
Gen. Tech.	0%	0%	55%	50%
Ext. Tech.	0%	0%	20%	0%
Total	0%	0%	40%	29%

5.3 Test on video and audio formats

For this test 4 files have been used, three video and one audio. The test has been made against two protocols, one for video and one for audio.

There is both a general part and an extended technical part for video. For audio only an technical metadata test is made. The result is presented in the table below.

As can be seen Jhove doesn't manage to open any of the files used in the test. File identifier and Metadata Extractor Tool only manages to get information from the format tag for video. Diagram 5 and table 11 shows that only Exiftool can handle video and audio.

The program manages the tags for: format, picture size, profile and no of video streams for all video files. For other tags the result varies between 33 and 67 per cent.

The tags chroma, bit rate mode, profile and no video streams causes Exiftool most problem. For audio it was bit rate mode and resolution where Exiftool failed to find information. The program could handle all other tags.

	Exiftool	File identifier	Jhove	MetaExtTool
Container	33%	0%	0%	0%
Bit rate	67%	0%	0%	0%
No. of video streams	0%	0%	0%	0%
Not of audio streams	33%	0%	0%	0%
Format	100%	67%	0%	67%
Profile	0%	0%	0%	0%

Bit rate mode	0%	0%	0%	0%
Bit rate	33%	0%	0%	0%
Picture size	100%	0%	0%	0%
Aspect ratio	67%	0%	0%	0%
Frame rate	100%	0%	0%	0%
Standard	33%	0%	0%	0%
Chroma	0%	0%	0%	0%

5.4 Conclusion of the tests on specific formats

Exiftool shows the best results on the test on technical metadata for the tiff format. Here the program gives best result on the extended metadata as well as total on the seven files used in the test. It doesn't do so well on the simple text files as it failed to open a single file. In the final test of video and audio formats it again comes out on top.

Jhove gives a very good result on technical metadata on the tiff format. As in previous tests it had problems, only managing to open 5 of the 7 files, which is reflected in the score. In the test of technical metadata on text formats Jhove is the only program that manages both general and extended technical metadata. Among the other programs only one gave any general technical metadata, the others gave zero information. When it comes to video and audio Jhove didn't manage to read any of the files used in the test, giving zero metadata.

Metadata Extractor Tool extracts all possible general technical metadata for tiff, which is a good result. The result for the extended part is however not so good. Here the program only finds information about 5 out of 10 tags. The program only manages the general technical part for the test of simple text formats, giving zero result on the extended part. The result of the last test, video and audio, is not good; only 7% for video and 17% for audio.

File identifier has the lowest score in 2 of 3 tests: the technical test on tiff and technical test of text formats. For video and audio the result is not much better. The program has the lowest total score in the test under section 5.

6 Conclusions

The conclusions that can be made from the test are: as can be seen from section 4 it's easiest to get metadata from image formats, which is also reflected in chapter 5.1. All programs do a good job here, possibly because image formats contain a lot of metadata and that it's structured in a good way. Also for text formats the programs give a lot of metadata. It's hardest to extract metadata from video and audio formats as can be seen in chapter 4 and 5.3. It is probably necessary to use programs specially design for video and audio formats.

Exiftool is best

If we rank the programs based on the tests in this report Exiftool (v 7.30) comes out as number one. It gives good results in the tests in chapter 4 and can handle a large number of formats (can read 14 of 16 files). It also gives both technical and descriptive metadata in chapter 4.2.

In the test against specific video and audio formats it gives good results. For video and audio it's the only program that does a reasonable job. The only weak point is that it cannot handle some simple text files (.txt).

Jhove

Jhove v 1.1 scores well in the test on technical and descriptive metadata on some few formats (ch 4.3), the best result being for tiff and PDF/A. In the test of the tiff format (5.1) it is the only program able to extract information for all tags in the extended list. In the test in ch 4 Jhove comes out as number 3 according to diagrams 1 and 2. This reflects the fact that Jhove only handles a limited number of formats.

It's also a result of the fact that if Jhove cannot validate a file it gives very little, if any, metadata. This permeates all test, degrading the result. Eg when testing for technical metadata for tiff (ch 5.1) since the program only can open 5 of the 7 files. Jhove doesn't handle the video and audio formats tested, giving zero result.

Metadata Extractor Tool

Metadata Extractor Tool v 3.1GA is the best program for extracting general technical metadata specified in the test protocol used in ch 4. This is shown clearly in diagram 2. Metadata Extractor Tool can be an alternative for extracting general technical metadata. The program can handle a reasonable number of different formats (13 out of 16 files).

It is also the only program that can read an Open office file. The weak points are: extracting descriptive metadata (ch 4) and extended technical metadata (ch 5).

File identifier

The only test where File identifier v 0.6.1 scored well was that it was able to read 14 out of 16 files (ch 4). The problem with this program is that it gives very little metadata, failing to satisfy the demands of the test protocols.

Appendices

Appendix 1: File lists and test files

Lists of the most common important file formats were produced by KB, SLBA and RA. They contain the most common format for each institution, see table below:

KB	SLBA	RA
text/html	Mpeg-1 layer3	tiff
jpeg	Mpeg-1 video	text/html
tiff	mpeg-2	jpeg
pdf	Motion jpeg-2k	pdf/a
gif	wave	Wave BWF

The 16 files used in the tests in ch 4 are divided into categories text, image and audio and video. In the category “other”, with subclasses tiff, text and video and audio, can be found the files used in ch 5. See table below:

File format	Version	
Text		
	Text utf-8	
	html	Hypertextmarkup L 4.01
	html	ExtensiveHypertextmarkup L 1.0
	Ms Word	
	Open Office text	1.0
	pdf/a-1b	1.4
	Pdf	1.3
Image		
	tiff	6.0
	tiff	6.0 profile, EXIF 2.2
	jpeg	1.01
	Jpeg	1.02
	Jpeg 2000	Part 2
	png	1.1
	gif	1989a
Audio/video		
	mpeg	Layer 3
	Mpeg-1	video

Other files		
Tiff tests	tiff	6.0
	Tiff (LZW)	6.0
	tiff	6.0
	tiff	6.0
	tiff	6.0
	tiff	6.0
	tiff	6.0 profile, EXIF 2.2
Text tests	text-ascii	
	text-ascii_pd	
	text-ebcdic	
	text-ebcdic	
	Text utf-8	
Video/audio	mpeg-2	video
	mpeg-4	video
	mpeg-1	Video format
	mpeg	
	mpeg	Layer 3

Appendix 2: KB list of general metadata

In this appendix we show the list of general metadata that has been used to test the programs for general technical and descriptive metadata. In total it has been used in 64 test protocols in ch 4.

List of general metadata from KB

Technical metadata			
Tag	Yes	No	Remark
format			
version			
Coding, eg BASE 64			
Character set			
size			
If truncated			
Checksum and algorithm			
Sum technical			
Descriptive metadata			
Title			
Identifier			
date			
creator/author			
publisher			
subject/keyword			
Sum descriptive			
Sum total			

Appendix 3

Here is shown the test protocol used for test against video formats. In total 12 protocols were completed in ch 5.3.

List of technical metadata from SLBA

General metadata			
tag	Yes	No	Remark
container			
Bit rate			
No of video streams			
No of audio streams			
Sum general metadata			
Technical metadata			
format			
profile			
Bit rate mode			
Bit rate			
Picture size			
Aspect ratio			
Frame rate			
standard			
chroma			
Sum technical metadata			
Sum total			

Appendix 4: Technical metadata for audio

This is the test protocol used when testing the programs against audio files. This protocol was used in in total 4 tests in ch 5.3.

General metadata			
Tag	Yes	No	Remark
format			
Bit rate mode			
Bit rate			
no. of channels			
Sample rate			
resolution			
Sum			

Appendix 5: technical metadata for tiff

This is the test protocol used when testing against tiff files. It has been used in the 28 tests in ch 5.1.

List of technical metadata for tiff from RA

Technical metadata			
tag	Yes	No	Remark
format			
version			
Coding, eg BASE 64			
Character set			
size			
Truncated?			
Checksum and algorithm			
Sum technical			
Extended techn. md			
NewSubfileType (254)			
SubfileType (255)			
BitsPerSample (258)			
Compression (259)			
PhotometricInterpretation (262)			
Thresholding (263)			
SamplesPerPixel (277)			
Xresolution (282)			
PlanarConfiguration (284)			
IPTC/NAA (33723)			
Sum ext. techn. md			
Sum total			



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>