Optima SC Inc.

http://www.optimasc.com

File identifier user manual

Field name	
Author	Carl Eric Codère
Document last modification date	2006-12-14
Document reference	MAN-S200401-02

Table of Contents

1 Overview

The file identifier software permits identifying files as well as getting information on directories from their content (thereafter called a resource), and not simply from their file extensions (in the case of files). Major features of the file identifier are as follow:

- Identified resources have their file extensions identified, as well as a descriptive comment indicating the type of resource.
- Optionally, identified resources can be categorized and their MIME media types identified. This permits getting statistics on the types of files stored on the system.
- Optionally, different methods for extracting metadata related to the resource are supplied, the currently supported methods are as follow:
    1. Embedded metadata for certain file types
    2. Embedded XMP packet information support [2]
    3. 4DOS/4NT/4OS2 descript.ion file support
    4. DEX (Description explorer) database support (compatible with descript.ion file)
    5. Sidecar XMP file support
    6. External PAD (Portable application description) files (the filename associated with resource is extracted from within the PAD file itself)
- The resource identifier database is a simple text file, and can be easily extended by end users.
- Extraction of the metadata to a DEX (Description Explore software) database is also supported. This permits to use DEX database compatible tools to edit and view the metadata.
- In certain cases (for the most common file types), full parsing of the file is done, and this permits detecting if the file format is valid or not.
- Optional calculation and export of CRC-32 values to an SFV (simple file verification) file for archival purposes.
- Optional generic HTML report files, with title, hyperlinks and comment fields, for easy navigation of the resources through a web browser.

2 Resource categories

All resources can be categorized to indicate the type of resource it is. The following categories currently exist:

- archive: A file that is usually compressed and that is usually composed of one or more files. This is the preferred format to use for exchanging files between computers. In a more general term, an archive can also be defined as being a directory or folder on a computer, since it contains other files.
- audio: Any file that is used exclusively for producing or reproducing sounds.

- code: Represents executable machine code, or code for a specific virtual machine.
- database: Any type of file that is used to represent some data in a structured fashion. This includes database, and spreadsheet formats.
- file system: Contains a disk image, or part of a disk image.
- font: Any type of file that is used to represent a graphical representation of a character or symbol.
- image: Any file that is a visual representation with limited or no animation and no audio components.
- metadata: Any file that is used exclusively as a source of information for other resources.
- model: File that represents a 3d model of one or more objects.
- palette: Represents resource that contain palette and color mappings.
- pim: File representing a contact list, a schedule or a personal to do list. (Personal information management file)
- text: Any type of file that is used to represent a text document, such as from a word processor.
- video: Any type of file that is used to represent an animation or video, either with or without audio components.

3 MIME identification

The MIME types that will be identified are only those which are registered through IANA, all other MIME types will be ignored, as they are non-official.

4 Metadata extraction

The file identifier permits different several extraction methods of files. The brief method (selected by the `-cb` command line option), simply identifies the resource, and gives information on the file extensions of the resource (if any). This is the fastest method to use the file identifier, and is the default extraction. It does not extract metadata, it only gives information on the file format.

The second method, selected by `-cs` (standard search) on the command line, checks for metadata in a standard way by first checking in the dex database, sidecar xmp files and then embedded within the resource itself, it also identifies the resource. It does not search for special embedded metadata standards though (such as embedded XMP).

The final and slowest method (up 2-3 times slower than the standard search method), also searches byte per byte in the file to see if special embedded metadata packets are in the resource, irrespective of the file format. It also verifies and parses any XML file in the directory where the resource is located to determine if there is PAD metadata associated with this information This is the surest way to make sure to extract all metadata that is supported by the software. This option is selected by the `-ch` command line option.

Metadata extraction (-cs or -ch) on complete file systems can take up to several hours on local storage, especially with big files.

5 Resource identifier database

The database is used each time the application is loaded, it contains the comments, as well as the MIME type and file extensions as well as how to identify the different resources. By default the the magic database file is searched in the same directory where the application was launched. The default database name is `magic.db`. It is possible to specify a different location for the database by using the `-m` command line option.

The database can easily be extended by hand to support identifying new resource types. Certain experience with the

magic file format is necessary on how to add new entries. For more information on the format of the magic database, consult the `Description Explorer magic database specification`. [1].

6 DEX Database export

The DEX database is a database that is compatible with 4DOS/4NT/4OS2 descript.ion files (see [3]), it is used by the `Description explorer` software package to store extracted and modified metadata for the different resources on disk. It is also used by JPSoft 4NT/4DOS to store the title of the different resources.

The file identifier permits exporting the extracted metadata to a DEX database, so that the metadata can then easily modified by the user with 4DOS/4NT or with `Description explorer`. To save the extracted metadata to the DEX database use the `-i` command line option. When this command is used, any existing DEX database information shall be overwritten from what is found in the resource.

7 Reports

7.1 HTML Reports

Reports to an html file is also available with the `-eh0` command line option. The generated report is created in each of the directories that were scanned and has the name `listing.htm`. It contains the files (as an hyperlink), as well as the extracted comments and file types (if known). It also contains an hyperlink to the origin of the files if they were extracted.

The report can be configured in a very sample way using stylesheets. When the report is generated it reads the stylesheet default.css in the directory where the executable is located and embeds it in the generated html. The generated report conforms to ISO HTML (ISO 15445).

The different classes that can be modified in default.css as well as their simple explanation is shown in the following table:

Field name	Description
`BODY`	Indicates the default style for the entire web page.
`P`	Gives the default type for paragraphs within the web page.
`table.main`	Gives the style information for the table that will contain all the information in the table.
`table.main th`	Gives the style information for the headings in the main table.
`table.main td`	Gives the style information for each of the cells in the main table
`table.main p.name`	Gives information on the style that shall be used to print out the resource name cell. The name cell text shall always be within <TT> and </TT> HTML elements.
`table.main p.type`	Gives information on the style that shall be used to print out the resource type cell.
`table.main p.title`	Gives information on the style that shall be used to print out the resource title cell.
`p.footer`	Indicate the style for the signature at the bottom of the web page.

7.2 CSV Reports (Professional version only)

Reports to an CSV file is also available with the command -ec line option. This option requires also the filename that will be used for the report. The old report will be overwritten if it already exists. The fields are described in the following table.

Field name	Description
File	Complete path and filename of the resource on disk
dc:title	This is the title of the resource.
dc:creator	An entity primarily responsible for making the content of the resource. An example is an artist of an MP3 file.
dc:subject	This is the subject of the resource, this is usually the same as keywords, and are usually separated by commas.
dc:date	A date associated with an event in the life cycle of the resource. The format of the date is in ISO 8601 format (YYYY-MM-DD). This date is taken internally from the resource and is not related to the filesystem dates.
dc:rights	Copyright statement of the resource
dc:source	Origin of the resource, for example an album name
dc:identifier	A unique identifier associated with this resource, such as an UUID or an ISBN number.
dc:contributor	Entities responsible for making contributions to the content of the resource. For example the composers of a song.
dc:publisher	An entity responsible for making the resource available. Examples of a Publisher include a person or an organization.
dc:coverage	The extent or scope of the content of the resource.
dc:relation	A reference to a related resource.
dc:type	The nature or genre of the content of the resource. For example the music type for audio files.
dex:source	Origin of this resource. This is usually an URL/URI where the resource was downloaded.
File comment	The actual file description associated with this file format
MIME Type	Registered MIME type for this file format.
File format identifier (FFID)	Registered file format identifier (FFID)
File Extensions	Usual file extensions for this file format

8 SFV file export

For resource integrity checking support, the application can also generate standard SFV files (using the standard CRC-32 algorithm). The name of the final sfv file is `check.sfv`, and it will be generated in the same directory where the data was processed. When this option is used, the -crc option is automatically enabled, since the CRC's of all resources must be calculated. This option is enabled with the -s command line option.

9 Command line options

This gives an overview and explanations of the different command line options (short option and long option)of the file identifier:

`-d`	`--debug`	Prints out some debug information
`-m [file]`	`--magic-file [file]`	Specifies an alternate name and path to the magic database. By default, the database is searched in the local directory with the name magic.db.
`-cb`	`--check-brief`	File identification only (no metadata). This is the default checking option.
`-cs`	`--check-standard`	Standard identification search
`-ch`	`--check-harder`	Extended identification search. This means to also search byte per byte in the file. Slowest, but most complete method of extracting metadata from resources.
	`--crc`	Calculate the CRC-32 of the file (.sfv compatible algorithm)
`-eh0`	`--report-html`	Create a simple HTML report of the found resources.
`-ec [file]`	`--report-csv [file]`	Create a global CSV report of the found resources with the specified report name (Professional version only)
`-s`	`--sfv`	Generate also an SFV file (automatically sets the --crc option).
`-i`	`--import`	Import the extracted metadata to the DEX database
`-r`	`--recursive`	Recurse into subdirectories
	`--help`	Show this message and exit
`-v`	`--version`	Print version and exit
	`--verbose`	Verbose mode, also prints skipped files and errors. Otherwise no information on errors is given at all.
`-cp [cp]`	`--codepage [cp]`	Indicates in what character encoding the output should be done in. By default, without this option, all text is output in ISO-8859-1 (Similar to codepage 1252 under Windows). Possible values of cp are: • ISO-8859-1: Selects ISO-8859-1 character encoding (the default). • UTF-8: Selects UTF-8 character encoding. • CP850: Selects MS-DOS multilangual character encoding.

10 Filename specifications

After the options come one or more file specifications. Each of these will be verified, and wildcards are accepted.

Hidden and system files are never searched.

11 Operating system specific behavior

All devices and system files will never be searched even if specified, and will be skipped, as they might cause problems in the software.

11.1 UNIX Systems

Since the shell automatically expands wilcards, the -r option will only work if the wildcard specifications are put in double quotes, such as "/var/*"

12 More information

To get the latest version of the file identifier, go to `http://www.magicdb.org`, it also contains information on different file formats, as well as tips on how to create new file formats.

The official web page for this software package is: `http://www.optimasc.com/products/fileid/`

You can get more information on our software products by contacting us at `info@optimasc.com`.

Carl Eric Codère, Optima SC Inc.

`October 2006`

13 Credits

Thanks to Mélanie Charbonneau for her help in designing the icon of this application.

14 Bugs and limitations

You can report bugs of the software on this site: `http://www.optimasc.com/bugs/` by selecting the File identifier (freeware) project. The above site also contains the current bugs of the software, as well as limitations of the software.

- If there is more than one file specification at the command-line, the `-sfv` and html reports options shall automatically be disabled. This should be fixed in the next release of the software.

15 Bibliography

[1] Description Explorer Magic Database, Optima SC Inc., Ref. no. SPC-S200401-01, 2004-10-08

[2] XMP Specification, Adobe Systems Incorporated, January 2004,

[3] 4DOS/4NT Description file extensions proposal, Optima SC Inc., SPC-S200401-00, 2004-09-14

[4] Portable application description specification, Association of shareware professionals, 2004