NAME

extract - determine meta-information about a file

SYNOPSIS

extract [**-abdfhLnrsvV**] [**-B** *language*][**-H** *hash-algorithm*][**-l** *library*][**-p** *type*]
    [**-x** *type*] *file ...*

DESCRIPTION

This manual page documents version 0.4.0 of the **extract** command.

extract tests each file specified in the argument list in an attempt to infer meta-information from it. Each file is subjected to the meta-data extraction libraries from **libextractor**.

libextractor classifies meta-information (also referred to as keywords) into types. A list of all types can be obtained with the **-L** option.

OPTIONS

-a
Do not remove any duplicates, even if the keywords match exactly and have the same type (i.e. because the same keyword was found by different extractor libraries).

-b
Display the output in BiBTeX format. This implies the **-d** option.

-B *LANG*
Use the generic plaintext extractor for the language with the 2-letter language code *LANG*. Supported languages are DA (Danish), DE (German), EN (English), ES (Spanish), IT (Italian) and NO (Norwegian).

-d
Remove duplicates only if the types match exactly. By default, duplicates are removed if the types match or if one of the types is *unknown* (in this case, the duplicate of unknown type is removed).

-f
add the filename(s) (without directory) to the list of keywords.

-h
Print a brief summary of the options.

-H *ALGORITHM*
Use the *ALGORITHM* to compute a hash of each file (possible algorithms are sha1 and md5).

-L
Print a list of all known keyword types.

-n
Do not use the default set of extractors (typically all standard extractors, currently mp3, ogg, jpg, gif, png, tiff, real, html, pdf and mime-types), use only the extractors specified with the **-l** option.

-r
Remove all duplicates disregarding differences in the keyword type.

-s
Split keywords at delimiters (space, comma, colon, etc.) and list split keywords to be of *unknown* type. This can also be done by loading the split-library. Using this option guarantees that the splitting is performed after all other libraries have been run. It is always performed before duplicate elimination.

-v
Print the version number and exit.

-V
Be verbose.

-B
Run the printable extractor (costly, generic extractor for binaries)

-l *libraries*
Use the specified *libraries* to extract keywords. The general format of libraries is *[[-LIBRARYNAME[:[-]LIBRARYNAME]*]* where *LIBRARYNAME* is a libextractor compatible library and typically of the form *libextractor_jpeg.so*. The minus before the libraryname indicates that this library should be run after all the libraries that were specified so far. If the minus is missing, the library is run before all previously specified libraries.

-p *type*
Print only the keywords matching the specified *type*. By default, all keywords that are found and not removed as duplicates are printed.

-x *type*
Exclude keywords of the specified *type* from the output. By default, all keywords that are found and not removed as duplicates are printed.

SEE ALSO

libextractor (3) - description of the libextractor library

EXAMPLES

```
$ extract test/test.jpg
comment - (C) 2001 by Christian Grothoff, using gimp 1.2 1
mimetype - image/jpeg

$ extract -Vf -x comment test/test.jpg
Keywords for file test/test.jpg:
mimetype - image/jpeg
filename - test.jpg

$ extract -p comment test/test.jpg
comment - (C) 2001 by Christian Grothoff, using gimp 1.2 1

$ extract -nV -l libextractor_png.so -p comment test/test.jpg test/test.png
Keywords for file test/test.jpg:
Keywords for file test/test.png:
comment - Testing keyword extraction
```

LEGAL NOTICE

libextractor and the extract tool are released under the GPL.

BUGS

A couple of file-formats (on the order of 10^3) are not recognized...

AUTHORS

extract was originally written by Christian Grothoff <christian@grothoff.org> and Vidyut Samanta <vids@cs.ucla.edu>. Use <libextractor@cs.purdue.edu> to contact the current maintainer(s).

AVAILABILITY

You can obtain the original author's latest version from http://ovmj.org/libextractor/.