File magic and libextract on Windows

July 16, 2007 / Gavin Jackson

cygwin programming windows

My current problem is quite simple, I have a directory full of files with no extensions. I would like to locate all word documents and run them through my header/footer extraction tool. Easy, I'll just use the file tool in cygwin to check the files binary "magic" pattern.

So I download the cygwin file package and start running some tests - initially all of my word, excel and powerpoint files were coming back as "Windows Installer" - so I modified the magic file - it then only reported the items as "Microsoft Office Documents" - it turns out that all Microsoft formats are essentially the same thing - Microsoft Ole Objects (that contain an internal filesystem structure).

So is there a way to perform a deeper analysis of the ole2 structure (to differentiate between the various different office formats)? Yes, there is a library available called libextrator - it provides an application called extract that allows you to "extract" metadata from files.

http://gnunet.org/libextractor/

I needed to use the windows build of this software, and it needed a bit of fiddling to get it to work. Step 1 - extract the archive Step 2 - copy all dlls in lib/libextract/*.dll to bin/ Step 3 - run extract using the following command: extract -l libextract_ole2 -f mydoc.doc

```
C:\Documents and Settings\Administrator\Desktop\blah\bin>e
ctor_ole2 -f mydoc.doc
filesize - 20.99 KB
filename - mydoc.doc
mimetype - application/msword
language - U.S. English
company - Vertex Systems Incorporated
paragraph count - 7
line count - 13
last saved by - Gav
character count - 147
template - Normal.dot
creation date - 2007-07-12T23:25:0
title - Text on Page 1 (Section 1)
word count - 42
page count - 7
creator - CNVT
date - 2007-07-13T00:15:0
generator - Microsoft Office Word
C:\Documents and Settings\Administrator\Desktop\blah\bin>
```

By looking at the file mime type, we can now determine the exact type of ole2 object we are dealing with.