

Tools for File Format Identification, Validation and Characterization

Bill Underwood

Georgia Tech Research Institute

Atlanta, Georgia, USA

CurateGear

Chapel Hill, NC

January 9, 2013

Motivation: Digital Curation Tools

Digital Curators need automated tools for:

- **Identification of file formats**
- **Validation of file formats, with pertinent error messages**
- **Extraction of metadata**
- **Viewing/playing/reading file formats**
- **Conversion of legacy formats to current/standard formats**

Motivation Digital Curation Tools

- **Identification: DROID/PRONOM; File/Magic**
- **Validation: Harvard's JSTOR/Harvard Object Validation Environment (JHOVE), UCDC (JHOVE2)**
- **Metadata Extractor: National Library of NZ Metadata Extractor; GNU libextractor**
- **Viewers/Players: NASAView, QuickView Plus, IrfanView, XNView, KeyView, Columbus viewer**
- **Conversion: XML Electronic Normalization of Archives (Xena), OpenOffice.org's Format Converter, Alchemy**

Definitions

- **A file format is a set of rules for encoding and decoding data or computer instructions in a file.**
- **A *file type* is a class of files with the same file format.**
- **A *file format signature* is invariant data in a file format that can be used to identify the file type (or format) of a file**

External File Format Identifiers

- **File Name Extensions**
- **Metadata stored in the operating system**
 - **MacOS HFS Creator Code & File Type Code**
 - **MacOS X Uniform Type Identifier (UTI)**
- **Multipurpose Internet Mail Extensions (MIME) media types**
- **PRONOM Persistent Universal Identifier (PUID)**

Linux File Command and Magic File

- **Unix (Linux) File Command and Magic File are probably the most widely used tool for file format identification.**
- **Magic number is the term used for the concept of an internal file format signature.**
- **The file command applies tests for magic numbers contained in the Magic file to files to determine their file type and relevant metadata.**

Some Limitations of the file Command and Magic File

- **Difficult to update the tests for magic numbers.**
- **Tests that may give conflicting results must be properly sequenced.**
- **Tests for magic numbers are not one-to-one with file types.**
- **Tests output metadata as well as file type.**
- **Tests for character set and language of text files needs refinement.**
- **Only a few tests for MS Windows file types.**
- **Tests for Magic numbers have not been rigorously tested**

Extensions of File Command and Magic File to overcome Limitations

- **File Format Library**
- **Magic for individual file formats**
- **Output of file command/magic file is File Format ID**
- **Rewriting file command code for identifying Characteristics of Text files and Document Types**
- **Defined approx. 1000 file format signatures**
- **Collected examples of approx. 800 of the file format types**
- **Created File Signature Database**
- **Verified that File Format Identifier with magic file correctly identifies approx. 800 File Types**

Georgia Tech File Format Library

GTRI File Format Listing

Home File Format Extensions Import Data

Page: 18 :: Pages: [1](#) - [10](#) | [11](#) , [12](#) , [13](#) , [14](#) , [15](#) , [16](#) , [17](#) , **18**

Sort by: FFL ID :: [Bottom of Page](#)

File Format Name Filter: [All](#) # [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#)

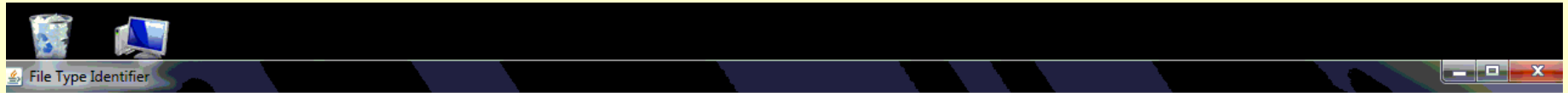
[Add New File Format](#)

| Row # | FFL ID | Name | Version | MIME | PUID |
|-------|--------|---------------------------------------------------------|---------|-----------------------------------------|-----------|
| 1 | 852 | AutoCAD Plotter Model Parameters File | | application/x-autocad-pmp | |
| 2 | 853 | CATIA Material Description | 5 | application/x-catia-material; version=5 | x-fmt/438 |
| 3 | 854 | Computer Graphics Metafile (Clear Text) | 2 | image/cgm; encoding=ascii; version=2 | |
| 4 | 855 | Computer Graphics Metafile (Clear Text) | 3 | image/cgm; encoding=ascii; version=3 | x-fmt/142 |
| 5 | 856 | Computer Graphics Metafile (Clear Text) | 4 | image/cgm; encoding=ascii; version=4 | fmt/302 |
| 6 | 857 | Computer Graphics Metafile (Binary) | 2 | image/cgm; encoding=binary; version=2 | fmt/304 |
| 7 | 858 | Computer Graphics Metafile (Binary) | 3 | image.cgm; encoding=binary; version=3 | fmt/305 |
| 8 | 859 | Computer Graphic Metafile (Binary) | 4 | image/cgm; encoding=binary; version=4 | fmt/306 |
| 9 | 860 | ESRI Arc/View Shapefile Index | | application/x-esri-shx | fmt/277 |
| 10 | 861 | SIARD Archive | 2.0 | application/siard; version=2 | fmt/161 |
| 11 | 862 | FLAC (Free Lossless Audio Codec) | | audio/flac | fmt/279 |
| 12 | 863 | PalmDOC | | application/x-palm-doc | |
| 13 | 864 | Plucker Document | | application/x-plucker | |
| 14 | 865 | Paint Shop Pro Image | 9 | image/x-paintshoppro; version=9 | |
| 15 | 866 | CATIA Process Description | 5 | application/x-catia-process; version=5 | |
| 16 | 867 | ESRI ArcGIS Projection File | | application/x-esri-prj | |
| 17 | 868 | ESRI Grid Header File | | application/x-esri-grid-header | |
| 18 | 869 | Electronic Arts VP6 | | video/x-ea-vp6 | |

Library Entry for ESRI Shape File Format

[Previous](#) | [Next](#) :: [Return To Listing](#) :: [Documentation](#) | [Viewers/Players/Extractors](#) | [Examples](#)

| | |
|-----------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <input type="button" value="Add"/> | Identification |
| <input type="button" value="Update"/> | Internal ID: 860 |
| <input type="button" value="Delete"/> | Name: ESRI Arc/View Shapefile Index |
| <input type="button" value="Get XML"/> | Version: |
| | Developed By: |
| | Application: |
| | Operating System: |
| | Date Released: |
| | File Extensions: shx |
| | MIME: application/x-esri-shx |
| | PUID: fmt/277 |
| | Object Class: |
| | Application Class: GIS |
| | Family: |
| | Signature |
| | Signature Description: File header 100 bytes long. Bytes 0-3 big-endian 9994. Bytes 4-23 unused, all hex00. File length, at byte 28 version = 1000 (little-endian long). At byte 100, first record offset big-endian long 50. |
| | Magic: <pre>0 belong 9994 >4 string \x00\x00\x00\x00 >>8 string \x00\x00\x00\x00\x00\x00\x00\x00\x00\x00 >>>16 string \x00\x00\x00\x00\x00\x00\x00\x00\x00 >>>>28 lalong 1000 >>>>100 belong 50 ESRI Arc/View Shapefile Index</pre> |
| | Signature Source: |
| | Precedes Signature: Return to Top to Page |
| | Documentation |
| <input type="button" value="Add"/> | Citation: ESRI Shapefile Technical Description |
| <input type="button" value="Update"/> | Rights: |
| <input type="button" value="Delete"/> | Source: www.esri.com/library/whitepapers/pdfs/shapefile.pdf |
| <input type="button" value="Add File"/> | |



File Edit View Help

| Filename | FileType | MimeType | Extension | PUID |
|------------------------------------------------------------------------------------|----------------------------------|------------------------|--------------|-----------|
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SciFiLaser_S08SF.357.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SciFiWhoosh_S08SF.1684.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SciFiWhoosh_S08SF.1684.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SemitruckHorn_S08IN.866.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SlingshotShoot_S08FO.2353.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SplashBallDrop_S08WR.88.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 0\SuctionPlop_S08CT.214.wav | EBU Broadcast Wave Format Ver 0 | audio/x-bwf; version=0 | wav bwf | fmt/1 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 1\96000_30ND_4.wav | EBU Broadcast Wave Format Ver 1 | audio/x-bwf; version=1 | wav bwf | fmt/2 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 1\short1.wav | EBU Broadcast Wave Format Ver 1 | audio/x-bwf; version=1 | wav bwf | fmt/2 |
| C:\Users\wu4\Documents\FFSamples\audio\EBU-BWF\Ver 1\short2.wav | EBU Broadcast Wave Format Ver 1 | audio/x-bwf; version=1 | wav bwf | fmt/2 |
| C:\Users\wu4\Documents\FFSamples\audio\flac\1.flac | FLAC (Free Lossless Audio Codec) | | | |
| C:\Users\wu4\Documents\FFSamples\audio\flac\applaud00.flac | FLAC (Free Lossless Audio Codec) | | | |
| C:\Users\wu4\Documents\FFSamples\audio\flac\BlueEyesExcerpt.flac | FLAC (Free Lossless Audio Codec) | | | |
| C:\Users\wu4\Documents\FFSamples\audio\flac\dropouts.flac | FLAC (Free Lossless Audio Codec) | | | |
| C:\Users\wu4\Documents\FFSamples\audio\IFF-8svx\8svx.Welcome On Amiga | IFF 8-bit Sampled Voice | audio/x-IFF-8svx | iff | x-fmt/157 |
| C:\Users\wu4\Documents\FFSamples\audio\m4a\Web_2_Workshop_Web_2.mp4.m4a | Apple iTunes AAC Audio | audio/x-m4a | m4a | |
| C:\Users\wu4\Documents\FFSamples\audio\midi\Bass_sample.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\Bass_sample2.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\bluegrass.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\Drum_sample.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\Drum_sample2.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\MIDI_sample.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\midi.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\midi.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\midi\testsnd.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\mp2\midi.mid | MIDI Audio | audio/x-midi | midi mid rmi | x-fmt/230 |
| C:\Users\wu4\Documents\FFSamples\audio\mp2\sample.mp2 | MPEG Audio Layer II | audio/mpa; layer=2 | mpw mpa mp2 | fmt/198 |
| C:\Users\wu4\Documents\FFSamples\audio\mp2\voice2.mp2 | MPEG Audio Layer II | audio/mpa; layer=2 | mpw mpa mp2 | fmt/198 |
| C:\Users\wu4\Documents\FFSamples\audio\mp2\voice3.mp2 | MPEG Audio Layer II | audio/mpa; layer=2 | mpw mpa mp2 | fmt/198 |
| C:\Users\wu4\Documents\FFSamples\audio\mp3\lock_19.mp3 | MPEG Audio Layer III | audio/mpa; layer=3 | mp3 | fmt/134 |

Messages

Workstation

Information Technology and Telecommunications Laboratory

9:04 PM
5/5/2011

Additional Information

GTRI url: <http://perpos.gtri.gatech.edu>

W. Underwood. Grammar-based Recognition of Documentary Forms and Extraction of Metadata. *The International Journal of Digital Curation*, Vol 5, Issue 1, 2010.

www.ijdc.net/index.php/ijdc/article/view/152

W. Underwood. Grammar-based Specification and Parsing of Binary File formats. *International Digital Curation Journal*, 2012.

Magic Test for Broadcast Wave Format Ver 1

Signature

Signature BWAVE PCM 1: RIFF header, WAVE id, bext chunk, version 1, fmt chunk, data chunk. BWAVE
Description: MPEG 1: RIFF header, WAVE id, bext chunk, version 1, fmt chunk, fact chunk

Magic:

```
# BWAVE PCM 1
0      string RIFF
>8     string WAVE
>>12  string bext
>>>&350 leshort 1
>>>>&254      search/32000      fmt\  \x10\x00\x00\x00\x01\x00
>>>>>&14      search/32000      data   EBU Broadcast Wave Format Ver 1
# BWAVE MPEG 1
0      string RIFF
>8     string WAVE
>>12  string bext
>>>&350 leshort 1
>>>>&254      search/3200      fmt\  \x28\x00\x00\x00\x50\x00
>>>>>&0      search/1000      fact\x04\x00\x00\x00      EBU Broadcast Wave Format Ver 1
```

Signature
Source:

Precedes
Signature:

Research Motivation

- **Digital Curators need the capability to automatically identify file formats for**
 - **Identifying appropriate format validation tool**
 - **Determining appropriate viewer, player, reader, extractor**
 - **Identifying Password recovery and decryption tools**
 - **Identifying repair tool for damaged files**