File Format Identification and Validation Tools

Roy Lechich 2/2007
Integrated Library & Technology Systems
Yale University Library

File Types and Formats

With the increasing uses for software, many of them involving storage of different kinds of data, numerous types of computer files have emerged. In fact , it is virtually impossible to list every type of computer file, because:

- some file types are general while others are very specific
- some file types are widely used while others may exist only for the use of one organization or individual
- the specfications for some file types are proprietary while others are open
- new types are always being created.

A file type often has an associated explicit format specification -- a formal declaration of how information is to be encoded in a file to qualify as being of that "type". The more widely used and specific a file type is, the more likely that it will have a corresponding format specification, with its formal declaration existing as a "reference implementation" and usually maintained by a standards organization.

Key file use concepts:

- different types are typically created, accessed and modified by some set of programs or tools – this can range from very general (e.g. text files able to be read and written by operating system tools) to very specific (e.g. a jpeg 2000 requiring a very specific program to create or view the image).
- users of a specific file will do so using these appropriate programs/tools

File Type Identification vs. File Format Validation

Most people typically are not overly concerned about file formats and related issues, because **a)** they are happily able to create and use their data with their sets of applications and tools, and **b**) they do not need to think about the implications of the passage of time upon their data and their ability to continue to use it. When data is considered with an eye to the future, and more specifically from the point of view of preservation, though, some important concerns surface.

Perhaps the two most immediate questions regarding a file's type and format are:

- how can we tell a file's type? And,
- if we know its type, how can we be sure that it conforms to its format specification so that we know it is still useable?

The terms used to address these questions are, respectively, file type identification and file format validation. Because of a possible range of scenarios, including human error and imperfect software, neither file identification or format vaildation are as simple as one might think.

Other Format-Related Issues

As a brief aside for the sake of a larger perspective, besides file type identification and file format validation, some other other format-related issues include: [1]

- Characterization:"What are the salient properties of what I have?"
- Assessment: "Is what I have usable?"
- Intervention: "How can I turn what I have into what I want?"

File Type Identification

By File Extension

The easiest way to get an indication of a file's type is simply to look at the filename extension (if there is one) -- the part of the name after the last dot (".") . This tells us what type the file "purports to be". A file called "myFile.txt", for example, purports to be a text file, while "myFile.jpg" purports to be a jpeg file. The problem with this first level of identification is that file extensions are not enforced -- since "myFile.txt" could easily be renamed "myFile.jpg" either by mistake, or maliciously, or simply because the user wants his text files to have a "jpg" extension for her own reasons, then the file would have a misleading extension and therefore could not be identified this way. Identifying a file's type by its extension is often very useful for some purposes, but not sufficient for any critical application.

By File Structure

The more complex file types, as mentioned above, have a rigorous format specification associated with it. By examining a file's structure and comparing it with known format specifications it should be possible to determine a file's type. Media files typically have in common that they have at least some kind of header section, containing certain information -- or metadata -- about the file, such as its type, its creation date, followed by a series of data "chunks" containing the actual data specific to the particular content. There are currently several file type identification tools that look at the header section to determine its type.

Format Registries

The idea of a format registry is to provide us with a database of the "known format specifications" mentioned above. As mentioned above, there are several issues regarding file format besides identification and validation, which are beyond the scope here, but often having to do with the fact that new formats are constantly being adapted while older ones are being abandoned, so a format registry also tries to contain helpful information regarding relationships between different formats, different format versions and specific migration information in order to help in the preservation of digital material. But in terms of the current discussion, a format registry aids in file type identification via examination of file structure.

File Format Validation

What is a Valid File?

Does a file's validity have to do with:

a) is it useable by the program or programs it is meant to be used with? or
b) is it fully compliant with its format specification?

Ideally, the answer to both these questions should always be the same. In reality though, programs can be imperfectly compliant with the format specification, since format specifications can be very complex, with very small and specific changes from one version to the next. The job of a file format validation tool is to read through the entire file, identifying and mapping each pertinent section to the specification, and determine the degree of compliance to that specification, and report the results, ideally including information about any variances. Yale University Library’s Rescue Repository service uses JHOVE (from Harvard / JSTOR – see *TOOLS* section below) during its ingest process to validate files.

File Format-Related Tools

Some of the currently available open source academic tools related to file format are:

Format Registries:

- **PRONOM:** developed by the Digital Preservation Department of the National Archives of the United Kingdom. PRONOM is a web-based format registry, designed to support digital preservation services. Services include the ability to search format information by format name and search software by formats handled. Core services are expected to be exposed as web services in a future version. PRONOM now uses a "PUID" (Persistent Unque Identifier) scheme to provide unique and persistent identifiers for records in its database.[2]

- Proposed / In Process
  - **Global Digital Format Registry (GDFR)**: a joint effort between Harvard University Library and OCLC with funding from the Mellon Foundation. This registry "will provide sustainable distributed services to store, discover, and deliver representation information about digital formats."[3] It is envisioned that PRONOM may end up as a node within this distributed system.
  - **Digital Formats for Library of Congress Collections**: This Library of Congress website is "collecting technical information about file formats relevant to the Library's digital collections, in order to inform preservation decisions. It also includes an overview of factors which may affect the sustainability of formats over the long-term".[4]

File Type Identification Tool:

- **DROID (Digital Record Object Identification)**: a software tool, also developed by the Digital Preservation Department of the National Archives of the United Kingdom, to perform automated batch file format identification, using the PRONOM registry. [5]

File Metadata Extract:

- **National Library of New Zealand Preservation Metadata Extract Tool:** similar in how it works to file type identification, this is a tool which extracts metadata from file headers. Like the extensible module-based design of JHOVE (below) , this Java tool uses "adapters" to extract metadata from filetypes including: MS Word, Word Perfect, Open Office, MS Works, MS Excel, MS PowerPoint, TIFF, JPEG, WAV, MP3, HTML, PDF,GIF, and BMP. This data is output in a standard XML format, allowing it to be uploaded into a preservation metadata repository. While incorporating JHOVE into a repository ingest process is useful as a validation step, the incorporation of this tool could also be valuable in an ingest process.[6]

File Type Identification and Validation Tool:

- **JHOV**E: this is a successful and widely used (inluding YUL) tool, developed by Harvard University Library and JSTOR, which does both file type identification and validation, as well as characterization. The format types which JHOVE can currently handle include: AIFF, ASCII, BYTESTREAM (as the default handling module when no other match is found), GIF, HTML, JPEG,JPEG2000, PDF, TIFF, UTF8, WAV and XML.
  JHOVE is configurable in many respects, including the ability to turn off full validation by specifying "short" mode, in which only the header's signature is analyzed, ability to include or exclude checksums in the output, and to choose from various output formats, including plain text and XML..

  Since JHOVE can do both file type identification as well as validation, it is currently Yale University Library's format-related tool of choice.

[1] from:
Stephen Abrams
Harvard University Library
DCC/LUCAS Joint Workshop presentation
Liverpool, 30 November-1 December, 2006
"Knowing What You've Got
Format Identification, Validation, and Characterization"

[2] PRONOM website: http://www.nationalarchives.gov.uk/pronom/

[3] from GDFR website: https://collaborate.oclc.org/wiki/gdfr/about.html

[4] from website: http://www.digitalpreservation.gov/formats/fdd/fdd000075.shtml

[5] DROID website: http://droid.sourceforge.net/wiki/index.php/Introduction

[6] NLNZ Extract Tool website: http://www.natlib.govt.nz/en/whatsnew/4initiatives.html#extraction