Follow-Up Questions

ASERL Webinar: "Intro to Digital Preservation #4 – Using FITS to Identify File Formats and Extract Metadata"

Speaker = Andrea Goethals, Harvard University
Session Recording: https://vimeo.com/65186522
Speaker's PPT: http://bit.ly/11yhVC8

1. **Any recommendations for tools that can validate sound & video files (and hopefully extract metadata, too)?**
   I don't know of very many tools that can validate file formats, presumably because identification and metadata extraction functions are more broadly needed than validation (especially for domains outside of preservation). Both JHOVE (http://jhove.sourceforge.net/index.html) and JHOVE2 (https://bitbucket.org/jhove2/main/wiki/Home) can validate WAVE audio, and JHOVE can also validate AIFF. MediaInfo (http://mediainfo.sourceforge.net/en) can identify and extract metadata for many different audio and video formats, but it doesn't validate formats.

2. **Any recommendations for basic tools to preserve metadata? (Coming from a small public library seeking to preserve MARC data)**
   Your options will depend on what form your metadata is in (standalone files, database, etc.) and what you need to do be able to do with it (search it, update it, etc.). Whatever you do use, make sure you have multiple copies of the metadata, and make sure that you can export it easily (so that it could be imported into newer technologies in the future).

3. **Could FITS launch a virus inside a file as part of its validation routine?**
   A file needs to be executed in order to launch any viruses it contains. FITS doesn't execute the files it parses - it scans their bits (similar to how antivirus software can parse files looking for viruses without launching viruses).

4. **Can FITS extract and pass on existing metadata that is embedded within a file (e.g., XMP embedded in an image file)?**
   Yes several of the tools used by FITS can extract metadata that is embedded in files. For example ExifTool can parse and extract XMP metadata as well as EXIF and other metadata.

5. **Does Harvard use a homegrown repository, or a paid service? Please provide a few details.**
   When we put our preservation repository into production in 2000 there weren't any open source or commercial repository software/services to use so we wrote our own. It's called the Digital Repository Service (DRS) - see http://hul.harvard.edu/ois/systems/drs/
   Internally it uses a lot of open source software (e.g. SOLR, JBOSS), as well as a few commercial (e.g. Oracle, Luratech Image Server).

6. **Does FITS include Jhove1 and Jhove2 in its basic package?**
   Just JHOVE (aka JHOVE1)

7. **Does PRONOM use a dynamic look-up, or is it a static data set that needs periodic updating**?
   The DROID tool ships with a signature file that contains format identification information from the PRONOM registry. You can update your DROID signature file separately - see http://digital-preservation.github.io/droid/

8. **I have found that Photoshop will over-write the metadata from original files (for Scanner Software Name, Scanner manufacturer, and Scanner Model Name). Do you know of a work-around?**
   Sorry, no.

9. **If you use FITS for large numbers of files (1000+ files), does FITS use a standard naming convention to keep track of the files it has processed?**
   When you input a single file to FITS the user must specify the name of the output file. When you input a directory of files to FITS, FITS writes the output files with the same names as the input files, appending '.fits.xml'.

10. **Can you recommend any tools for querying the FITS XML output?**
    The only tool I know of that is specific to the FITS XML output is C3PO (http://ifs.tuwien.ac.at/imp/c3po). It will take in FITS output and help you analyze content. Otherwise, you can use any generic XML parsers with FITS.

11. **Is there a registry that preserves/describes that actual format specifications themselves? (Is that what PRONOM does)?**
    Both the PRONOM (http://www.nationalarchives.gov.uk/PRONOM/Default.aspx) and UDFR (http://udfr.org/) format registries have the ability to describe format specifications as well as formats. For example see:
    http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=617&strPageToDisplay=documentation
    But not all formats listed in these registries have their specifications documented in these registries.

12. **Can you provide examples of interfaces to connect to the FITS API?**
    The FITS API is a Java API meant to be used within Java programs. For more about the API see https://code.google.com/p/fits/wiki/developer_documentation

13. **Please explain why you think using the API is better than using the command line.**
    It's faster. When you start FITS from the command-line your Java Virtual Machine (JVM) has to start up every single time you run it from the command-line. When you use a program that uses the FITS java API, your JVM only starts up once when your program starts.