FITS – The File Information Tool Set
Poster Proposal – Open Repositories 09
Randy Stern, Spencer McEwen -- Harvard University Library

Background: In order to adequately perform a preservation function, preservation digital repositories require sufficient technical metadata about files to perform the analyses required to identify and take preservation actions for at-risk file formats. Some digital repositories limit ingest and storage of digital files to formats that are well characterized and described by submitters. However, the requirement to provide technical metadata can be an inhibitor to deposit, thus limiting and preventing the deposit of large classes of materials. This problem can be mitigated by an automated file format identification and characterization process that performs as good a job as possible at extracting useful technical metadata on arbitrary files.

FITS: Harvard University Library has developed the File Information Tool Set (FITS), an open source java program that identifies, validates and extracts technical metadata for a wide range of file formats. It is designed to combine the strengths of any number of file analysis tools into a single unified file characterization that can subsequently be mapped to standard metadata formats, such as MIX, TextMD, etc. It acts as a wrapper, invoking and managing the output from several other open source tools. Output from these tools are converted into a common format, compared to one another and consolidated into a single XML output file. FITS is written in Java and is compatible with Java 1.5 or higher. The external tools currently used are:

* [http://hul.harvard.edu/jhove/ Jhove]
* [http://www.sno.phy.queensu.ca/~phil/exiftool/ Exiftool]
* [http://meta-extractor.sourceforge.net/ National Library of New Zealand Metadata Extractor]
* [http://droid.sourceforge.net/ DROID]
* [http://schmidt.devlib.org/ffident/index.html FFIdent]
* [http://unixhelp.ed.ac.uk/CGI/man-cgi?file File Utility]

FITS will be released as open source, and can easily be integrated into preservation workflows. FITS can be used as a command line tool or within other projects using its java API.

FITS is designed to be extensible. Any type of tool, whether it's based on Perl, Java, or something else entirely, can be added to FITS. A tool wrapper is created that encapsulates the complexities of invoking the tool, capturing the output, and converting it to FITS XML. For example Exiftool is written in Perl. The Exiftool tool wrapper checks for the operating system type and if Perl is installed. It then can decide if it should use the standard Perl version of Exiftool or the windows executable. It is the responsibility of the tool wrapper to convert the tool output into FITS XML and return a valid ToolOutput object. For tools that natively return XML, XSLT can be used to convert the output to FITS XML. For tools that do not return XML, the output can either be a) directly converted to FITS XML, or b) converted to a basic intermediate XML format and then converted using XSLT.