Using FITS to Identify File Formats and Extract Metadata

Andrea Goethals, Harvard Library

ASERL Webinar 2013

File Information Tool Set

What I'll cover

- Intro to…
  - File formats
  - File tools
- FITS

“File Format”

- “Specific structure or arrangement of data code stored as a computer file. A file format tells the computer how to display, print, and process, and save the data.”
  - BusinessDictionary.com
- “The organization of information according to preset specifications”
  - TheFreeDictionary

JFIF.TXT - Notepad

File Edit Format View Help

JPEG File Interchange Format
Version 1.02

September 1, 1992

Eric Hamilton
C-Cube Microsystems
1778 McCarthy Blvd.
Milpitas, CA 95035

+1 408 944-6300
Fax: +1 408 944-6314
E-mail: eric@c-cube.com

JPEG File Interchange Format
Version 1.02

Why a File Interchange Format

JPEG File Interchange Format is a minimal file format which enables JPEG bitstreams to be exchanged between a wide variety of platforms and applications. This minimal format does not include any of the advanced features found in the TIFF JPEG specification or any application specific file format. Nor should it, for the only purpose of this simplified format is to allow the exchange of JPEG compressed images.

JPEG File Interchange Format features

- Uses JPEG compression
- Uses JPEG interchange format compressed image representation
- PC or Mac or Unix workstation compatible
- Standard color space: one or three components. For three components, YCbCr (CCIR 601-256 levels)
- APP0 marker used to specify Units, X pixel density, Y pixel density, thumbnail
- APP0 marker also used to specify JFIF extensions
- APP0 marker also used to specify application-specific information

Macintosh WordPerfect 4.0

File Format Manual

A Guide to

Understanding and Interfacing to

Macintosh WordPerfect 4.0

January 4, 2011

INTERNATIONAL STANDARD

ISO/IEC 15444-1

First edition
2000-12-15

Information technology — JPEG 2000 image coding system —

Part 1: Core coding system

Technologies de l'information — Système de codage d'image JPEG 2000 —

Partie 1: Système de codage de noyau

INTERNATIONAL STANDARD

ISO/IEC 15444-1

First edition
2000-12-15

Annex I

JP2 file format syntax

(This Annex forms a normative and integral part of this Recommendation | International Standard. This Annex is optional for the minimum decoder.)

In this Annex and all of its subclauses, the flow charts and tables are normative only in the sense that they are defining an output that alternative implementations shall duplicate.

I.1 File format scope

Even so …

- Unclear specifications
- Complex/long specifications
- Specifications that depend on many other specifications

or

- No accessible sources (proprietary formats, very old formats)

Further complications for tool builders and users

- Related formats, examples:
  - OpenDocument formats are packaged as ZIP files
  - Many formats (XML, HTML, PS) are text formats
- Some formats lack obvious identifying features (e.g. magic numbers), examples:
  - Text character encoding
  - TIFF versions

Implications for file tools

- Can be hard for tools to accurately identify formats
- Some tools are more specific than others for particular formats
  - E.g. Zip vs. OpenDocument vs. OpenDocument Spreadsheet
- Some subjectivity behind format tools
  - Different names for same format
  - Different opinions about format validity

File tools

- Identify formats
- Validate formats
- Extract metadata

Poll 1

	Identify formats	Validate formats	Extract metadata	Formats
DROID	YES	NO	NO	> 1000
ExifTool	YES	NO	YES	couple hundred
FFident	YES	NO	NO	~ 50
File utility	YES	NO	NO	> 1000
JHOVE	YES	YES	YES	11 + variations
MediaInfo	YES	NO	YES	~30 A/V containers
NLNZ ME	YES	NO	YES	~ 20

Why FITS?

- Original motivation
  - Offset risk of accepting **any** format (Web archives, email attachments, donated hard drives)
- Thoughts
  - No single format identification tool can suffice (format support varies, accuracy varies)
  - Unsustainable to only use “library” tools – want to incorporate tools from any domain

Polls 2 & 3

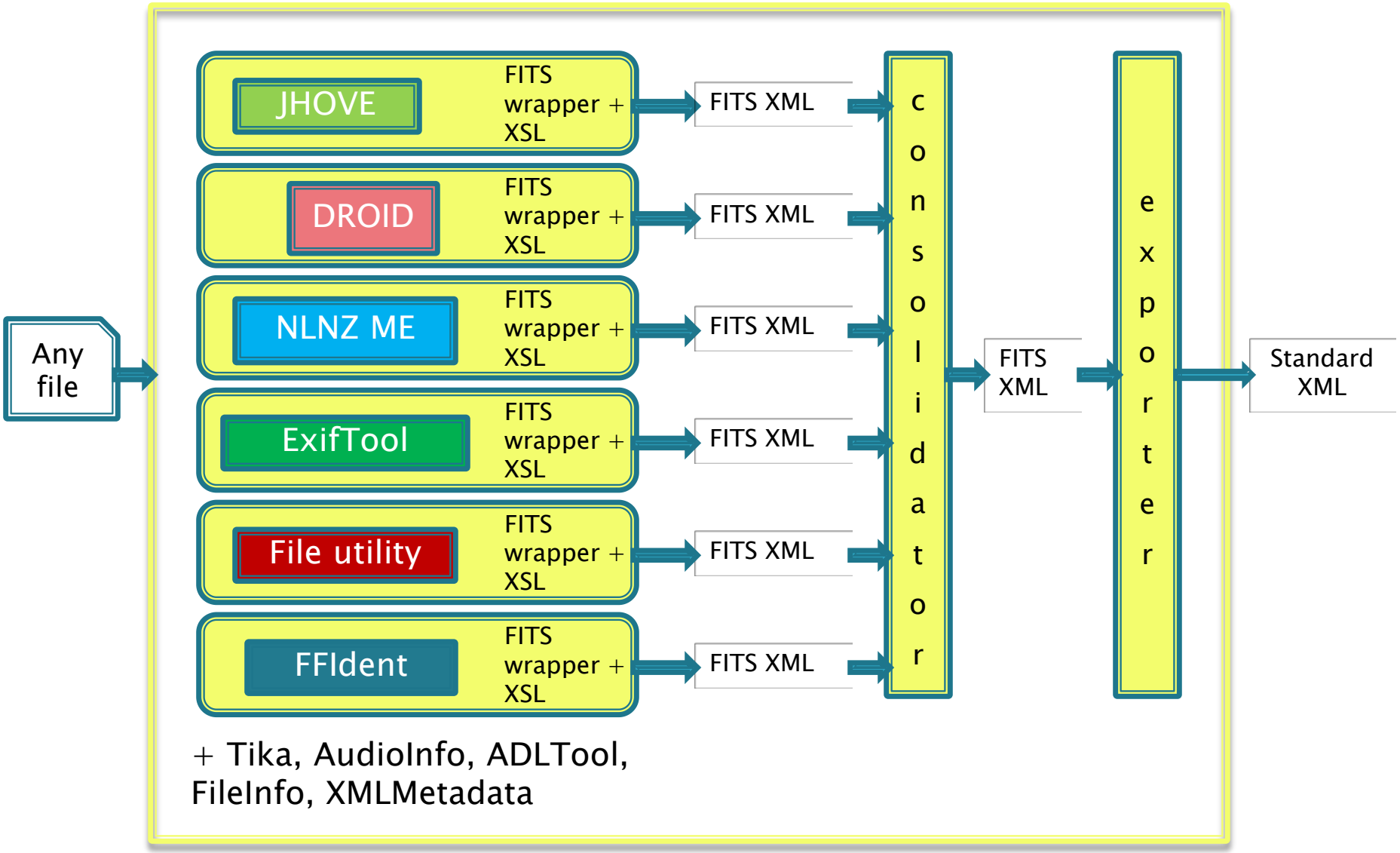
FITS strategy

- Develop a tool manager instead of a tool
- Include open source tools from any domain
- Make highly configurable, tweak over time as experience & knowledge is gained
- Account for tool inaccuracy in the design
  - Check the tools against each other
    - Do any disagree?
    - How many are in agreement?

What does it do?

- Identify many file formats
- Validate a few file formats
- Extract metadata
- Calculate basic file info (file size, MD5, etc.)
- Output technical metadata
  - Community-standard metadata schemas
- Identify problem files
  - Conflicting opinions on format, metadata values
  - Unidentifiable file formats

The process


Any file
JHOVE
DROID
NLNZ ME
ExifTool
File utility
FFIdent
FITS wrapper + XSL
FITS XML
consolidator
FITS XML
exporter
Standard XML
+ Tika, AudioInfo, ADLTool, FileInfo, XMLMetadata

Normalization

- Different names for the same format
  - ‘JPEG2000’ vs ‘JPEG 2000’ vs ‘JPEG 2000 image”
- Different values for the same metadata
  - “inches” vs “2” vs “in.”
  - “Grayscale” vs “Greyscale”
- Different ways of saying it can’t identify it
  - ‘Unknown Binary’ vs ‘bytestream’ vs ‘data’ vs no value
  - ‘application/octet-stream’ vs ‘application/unknown’ vs no value
- Different ways metadata is output
  - Ex: bits per sample (single or multiple values)

Fits output

```
<fits>
        <identification>
                // format name, version, registry IDs
        </identification>
        <fileinfo>
                // file name, size, MD5, etc.
        </fileinfo>
        <filestatus>
                // validity info
        </filestatus>
        <metadata>
                // normalized, combined metadata
        </metadata>
        <toolOutput>
                // native tool output
        </toolOutput>
</fits>
```

HARVARD LIBRARY

Format identification metadata

- Format
  - Name
  - Version
- MIME media type
- Format registry identifier(s)
  - PRONOM puid

Format identification metadata

- Format
  - Name = Portable Document Format
  - Version =1.4
- MIME media type = application/pdf
- Format registry identifier(s)
  - PRONOM puid = fmt/16

Demos: basic command line

cmd (open up a shell)

cd ..\..\Program Files\Fits\fits-0.6.2 (navigate to install)

.\fits.bat -h (see parameters)

.\fits.bat -i ..\testfiles\myfile.pdf (FITS XML metadata only)

<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.6.2" timestamp="4/24/13 10:40 AM">
  <identification>
    <identity format="Portable Document Format" mimetype="application/pdf" toolname="FITS" toolversion="0.6.2">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Exiftool" toolversion="9.06" />
      <tool toolname="Droid" toolversion="3.0" />
      <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
      <tool toolname="ffident" toolversion="0.2" />
      <version toolname="Jhove" toolversion="1.5">1.6</version>
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/20</externalIdentifier>
    </identity>
  </identification>
  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">12613</size>
    <creatingApplicationName toolname="NLNZ Metadata Extractor" toolversion="3.4GA" status="SINGLE_RESULT">/</creatingApplicationName>
    <lastmodified toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:40:05-04:00</lastmodified>
    <created toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:39:31-04:00</created>
    <filepath toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">C:\Program Files\Fits\fits-0.6.2\..\testfiles\myfile.pdf</filep
ath>
    <filename toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">..\testfiles\myfile.pdf</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">40d6569758762af9ff5c6046b6a1ad2f</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">1366814405195</fslastmodified>
  </fileinfo>
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</valid>
  </filestatus>
  <metadata>
    <document>

timestamp="4/24/13 10:40 AM">
  <identification>
    <identity format="Portable Document Format" mimetype="application/pdf" toolname="FITS" toolversion="0.6.2">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Exiftool" toolversion="9.06" />
      <tool toolname="Droid" toolversion="3.0" />
      <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
      <tool toolname="ffident" toolversion="0.2" />
      <version toolname="Jhove" toolversion="1.5">1.6</version>
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/20</externalIdentifier>
    </identity>
  </identification>
  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">12613</size>
    <creatingApplicationName toolname="NLNZ Metadata Extractor" toolversion="3.4GA"
status="SINGLE_RESULT">/</creatingApplicationName>
    <lastmodified toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24
10:40:05-04:00</lastmodified>
    <created toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:39:31-
04:00</created>
    <filepath toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">C:\Program
Files\Fits\fits-0.6.2\..\testfiles\myfile.pdf</filepath>
    <filename toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">..\testfiles\myfile.pdf</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">40d6569758762af9ff5c6046b6a1ad2f</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">1366814405195</fslastmodified>
  </fileinfo>
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</valid>
  </filestatus>
  <metadata>
    <document>
      <title toolname="Jhove" toolversion="1.5" status="CONFLICT">Local Disk</title>

<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.6.2" timestamp="4/24/13 10:40 AM">
  <identification>
    <identity format="Portable Document Format" mimetype="application/pdf" toolname="FITS" toolversion="0.6.2">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Exiftool" toolversion="9.06" />
      <tool toolname="Droid" toolversion="3.0" />
      <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
      <tool toolname="ffident" toolversion="0.2" />
      <version toolname="Jhove" toolversion="1.5">1.6</version>
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/20</externalIdentifier>
    </identity>
  </identification>
  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">12613</size>
    <creatingApplicationName toolname="NLNZ Metadata Extractor" toolversion="3.4GA" status="SINGLE_RESULT">/</creatingApplicationName>
    <lastmodified toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:40:05-04:00</lastmodified>
    <created toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:39:31-04:00</created>
    <filepath toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">C:\Program Files\Fits\fits-0.6.2\..\testfiles\myfile.pdf</filep
ath>
    <filename toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">..\testfiles\myfile.pdf</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">40d6569758762af9ff5c6046b6a1ad2f</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">1366814405195</fslastmodified>
  </fileinfo>
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</valid>
  </filestatus>
  <metadata>
    <document>
      <title toolname="Jhove" toolversion="1.5" status="CONFLICT">Local Disk</title>
      <title toolname="Exiftool" toolversion="9.06" status="CONFLICT">C:\Program Files\Fits\testfiles\Felix_output.txt</title>
      <pageCount toolname="Jhove" toolversion="1.5">2</pageCount>

```
    <lastmodified toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:40:05-04:00</lastmodified>
    <created toolname="Exiftool" toolversion="9.06" status="SINGLE_RESULT">2013:04:24 10:39:31-04:00</created>
    <filepath toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">C:\Program Files\Fits\fits-
0.6.2\..\testfiles\myfile.pdf</filep
ath>
    <filename toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">..\testfiles\myfile.pdf</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">40d6569758762af9ff5c6046b6a1ad2f</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">1366814405195</fslastmodified>
  </fileinfo>
  <filestatus>
    <well-formed toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</well-formed>
    <valid toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">true</valid>
  </filestatus>
  <metadata>
    <document>
      <title toolname="Jhove" toolversion="1.5" status="CONFLICT">Local Disk</title>
      <title toolname="Exiftool" toolversion="9.06" status="CONFLICT">C:\Program
Files\Fits\testfiles\Felix_output.txt</title>
      <pageCount toolname="Jhove" toolversion="1.5">2</pageCount>
      <isTagged toolname="Jhove" toolversion="1.5">no</isTagged>
      <hasOutline toolname="Jhove" toolversion="1.5" status="CONFLICT">yes</hasOutline>
      <hasOutline toolname="NLNZ Metadata Extractor" toolversion="3.4GA"
status="CONFLICT">no</hasOutline>
      <hasAnnotations toolname="Jhove" toolversion="1.5"
status="SINGLE_RESULT">no</hasAnnotations>
      <isRightsManaged toolname="Exiftool" toolversion="9.06"
status="SINGLE_RESULT">no</isRightsManaged>
      <isProtected toolname="Exiftool" toolversion="9.06">no</isProtected>
      <hasForms toolname="NLNZ Metadata Extractor" toolversion="3.4GA"
status="SINGLE_RESULT">no</hasForms>
    </document>
  </metadata>
</fits>
```

Format-specific technical metadata

- For text: TextMD (Library of Congress)
- For images: MIX (Library of Congress)
- For documents: DocumentMD (Florida Virtual Campus / Harvard Library)
- For audio: AES57 (Audio Engineering Society)

Demos: extended metadata

.\fits.bat –i ..\testfiles\myfile.pdf **–xc** (FITS XML metadata+ standard technical metadata)

<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.6.2" timestamp="4/24/13 10:48 AM">
  <identification>
    <identity format="Portable Document Format" mimetype="application/pdf" toolname="FITS" toolversion="0.6.2">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Exiftool" toolversion="9.06" />
      <tool toolname="Droid" toolversion="3.0" />
      <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
      <tool toolname="ffident" toolversion="0.2" />
      <version toolname="Jhove" toolversion="1.5">1.6</version>
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/20</externalIdentifier>
    </identity>
  </identification>
                .
                .
                (snip)
                .
                .
    <standard>
      <docmd:document xmlns:docmd="http://www.fcla.edu/docmd">
        <docmd:PageCount>2</docmd:PageCount>
        <docmd:Features>hasOutline</docmd:Features>
       </docmd:document>
    </standard>
  </document>
  </metadata>
</fits>

In our AIPs

- 1000008.xml
  - Premis:
    - Generic technical metadata (fixity, size, format, creating application)
    - Format-specific technical metadata in objectCharacteristicsExtension
  - FITS XML output in administrative metadata

Very configurable

- xml/ directory
- fits.xml (tweak your tool preferences)
- fits_format_tree.xml (tweak knowledge-base of related formats)

```
<branch format="JPEG 2000">
    <branch format="JPEG 2000 JP2"/>
    <branch format="JPEG 2000 JPX"/>
</branch>
```

Project home

- http://fits.googlecode.com
  - Downloads: get the newest version
  - Mailing list: fits-users (new releases announced here)
  - Issues: File any bugs
- Source code (if you want to contribute): https://github.com/harvard-lts/fits

Feedback survey

- http://www.surveymonkey.com/s/2013_DP4

(If time) Conflict reports

C:\Program Files\Fits\fits-0.6.1>.\fits.bat -i demo\Acknowledgements.rtf

```
<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://hul.
harvard.edu/ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.6.1"
timestamp="7/21/12 3:51 PM">
  <identification status="CONFLICT">
    <identity format="Plain text" mimetype="text/plain" toolname="FITS" toolversion="0.6.1">
        <tool toolname="Jhove" toolversion="1.5" />
    </identity>
    <identity format="Rich Text Format" mimetype="application/rtf, text/rtf" toolname="FITS" toolversion="0.6.1">
        <tool toolname="Droid" toolversion="3.0" />
        <version toolname="Droid" toolversion="3.0" status="CONFLICT">1.5</version>
        <version toolname="Droid" toolversion="3.0" status="CONFLICT">1.6</version>
        <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/50</externalIdentifier>
        <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/51</externalIdentifier>
    </identity>
    <identity format="Rich Text Format" mimetype="text/rtf" toolname="FITS" toolversion="0.6.1">
        <tool toolname="ffident" toolversion="0.2" />
    </identity>
</identification>
```

HARVARD LIBRARY

(If time) Conflict reports

- Indicate tool inaccuracies and/or areas for educating ourselves
- To resolve
  - Is Rich Text Format a more specific form of Plain Text?
    - If so, adjust fits_format_tree.xml
  - What should the MIME media-type for Rich Text Format? (consult specification if possible)
    - Normalize the tool output to this MIME media-type