FITS Demo

Digital Preservation 2012

Andrea Goethals

File Information Tool Set

Why FITS? original motivation:

- Offset risk of accepting **any** format
  - Web archives, email attachments, opaque objects
- No single format identification tool can suffice (format support varies, accuracy varies
- Difficult to use multiple tools together (language differs)
- Unsustainable to only use “library” tools – want to incorporate tools from any domain

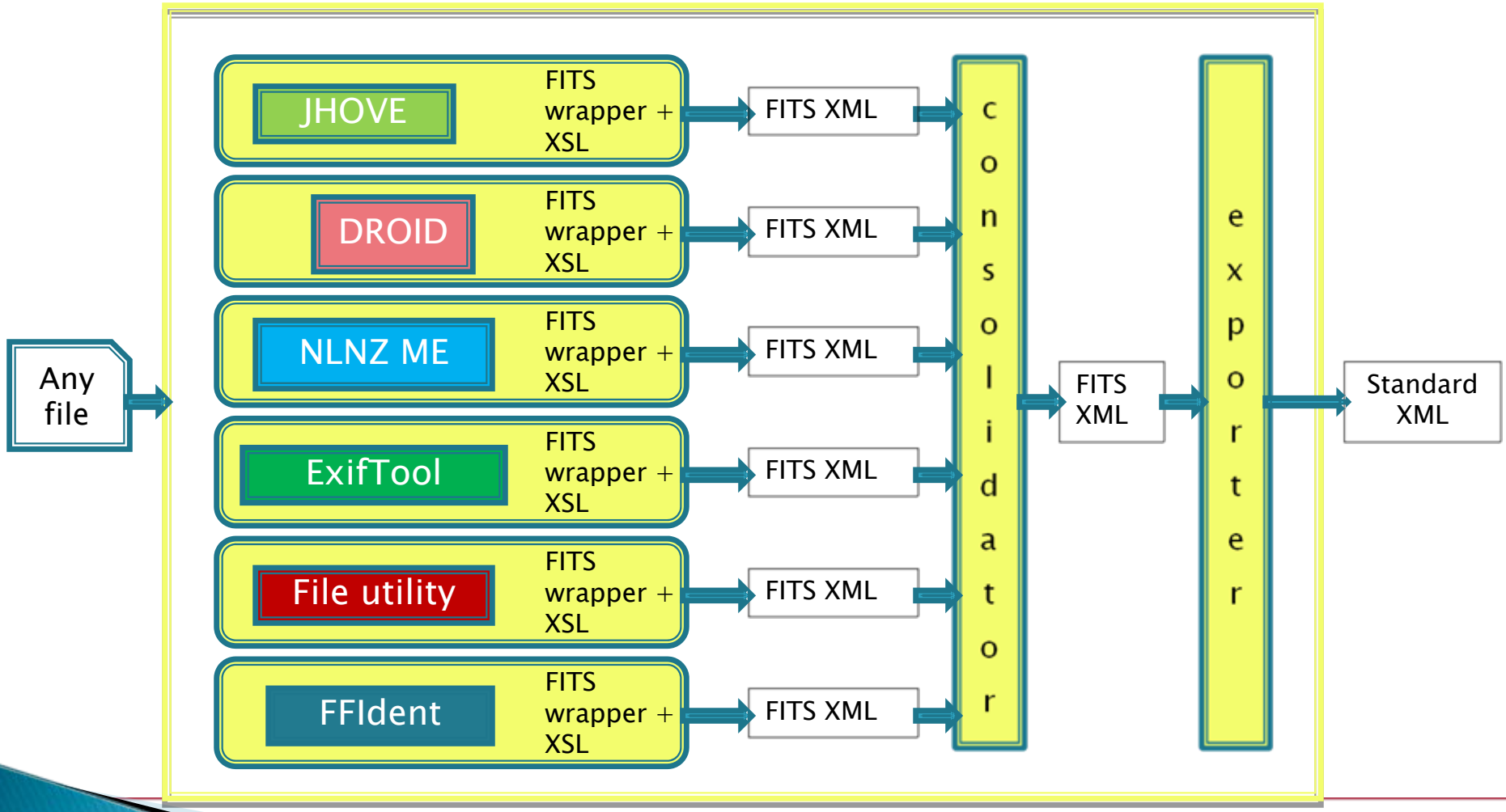
FITS Strategy

- 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 technical 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
  - Empty files
  - Technical metadata can't be generated

The process



HARVARD LIBRARY

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

Demos: basic command line

cmd (open up a shell)
..\..\Program Files\Fits\fits-0.6.1 (navigate to install)
.\fits.bat -h (see parameters)
.\fits.bat -i RELEASE.txt (FITS 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="0.6.1" timestamp="7/20/12 5:01 PM">
  <identification>
    <identity format="Plain text" mimetype="text/plain" toolname="FITS" toolversion="0.6.1">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Droid" toolversion="3.0" />
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">x-fmt/111</externalIdentifier>
    </identity>
  </identification>
  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">7838</size>
    <filepath toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">C:\Program Files\Fits\fits-0.6.1\RELEASE.txt</filepath>
    <filename toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">RELEASE.txt</filename>
    <md5checksum toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">7dc74a990c85006fa028ec8fbdbc0d20</md5checksum>
    <fslastmodified toolname="OIS File Information" toolversion="0.1" status="SINGLE_RESULT">1335359242000</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>
    <text>
      <linebreak toolname="Jhove" toolversion="1.5">CR/LF</linebreak>
      <charset toolname="Jhove" toolversion="1.5">US-ASCII</charset>
    </text>
  </metadata>
</fits>

HARVARD LIBRARY

Demos: basic command line

.\fits.bat –i RELEASE.txt –x (standard technical metadata only)

```
<?xml version="1.0" encoding="UTF-8"?>
   <textMD:textMD xmlns:textMD="info:lc/xmlns/textMD-v3"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="info:lc/xmlns/text
MD-v3 http://www.loc.gov/standards/textMD/textMD-v3.01a.xsd">
      <textMD:character_info>
        <textMD:charset>US-ASCII</textMD:charset>
        <textMD:linebreak>CR/LF</textMD:linebreak>
      </textMD:character_info>
   </textMD:textMD>
```

HARVARD LIBRARY

Demos: basic command line

.\fits.bat -i RELEASE.txt -xc (FITS metadata+ standard technical metadata)

```
<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="7/20/12 5:11 PM">
  <identification>
    <identity format="Plain text" mimetype="text/plain" toolname="FITS" toolversion="0.6.1">
      <tool toolname="Jhove" toolversion="1.5" />
      <tool toolname="file utility" toolversion="5.03" />
      <tool toolname="Droid" toolversion="3.0" />
      <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">x-fmt/111</externalIdentifier>
    </identity>
  </identification>
.
(snip)
.
  <metadata>
    <text>
      <linebreak toolname="Jhove" toolversion="1.5">CR/LF</linebreak>
      <charset toolname="Jhove" toolversion="1.5">US-ASCII</charset>
      <standard>
        <textMD:textMD xmlns:textMD="info:lc/xmlns/textMD-v3">
          <textMD:character_info>
            <textMD:charset>US-ASCII</textMD:charset>
            <textMD:linebreak>CR/LF</textMD:linebreak>
          </textMD:character_info>
        </textMD:textMD>
      </standard>
    </text>
  </metadata>
</fits>
```

HARVARD LIBRARY

Demos: basic command line

.\fits.bat –i RELEASE.txt –o demo\RELEASE_out1.txt
(FITS metadata only written to a file)

HARVARD LIBRARY

In our AIPs

- od_1000012.xml
  - premis:fixity (MD5)
  - premis:size (file size)
  - premis:format
  - premis:creatingApplication
  - premis:objectCharacteristicsExtension (documentMD)
  - hulDrsAdmin:fileIdentification
  - hulDrsAdmin:formatValidation
  - hulDrsAdmin:suppliedFilename
  - hulDrsAdmin:suppliedDirectory

Main configuration: fits.xml

- In fits-0.6.1/xml directory
- Key items
  - Enable/disable tools
  - Add new tools
  - Tools to prefer
  - Prevent tools from processing files by file extension
  - Option to include tools' native output
  - Report or ignore conflicts

Configuration: fits_format_tree.xml

- In fits-0.6.1/xml directory
- To indicate more specific formats

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

HARVARD LIBRARY

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

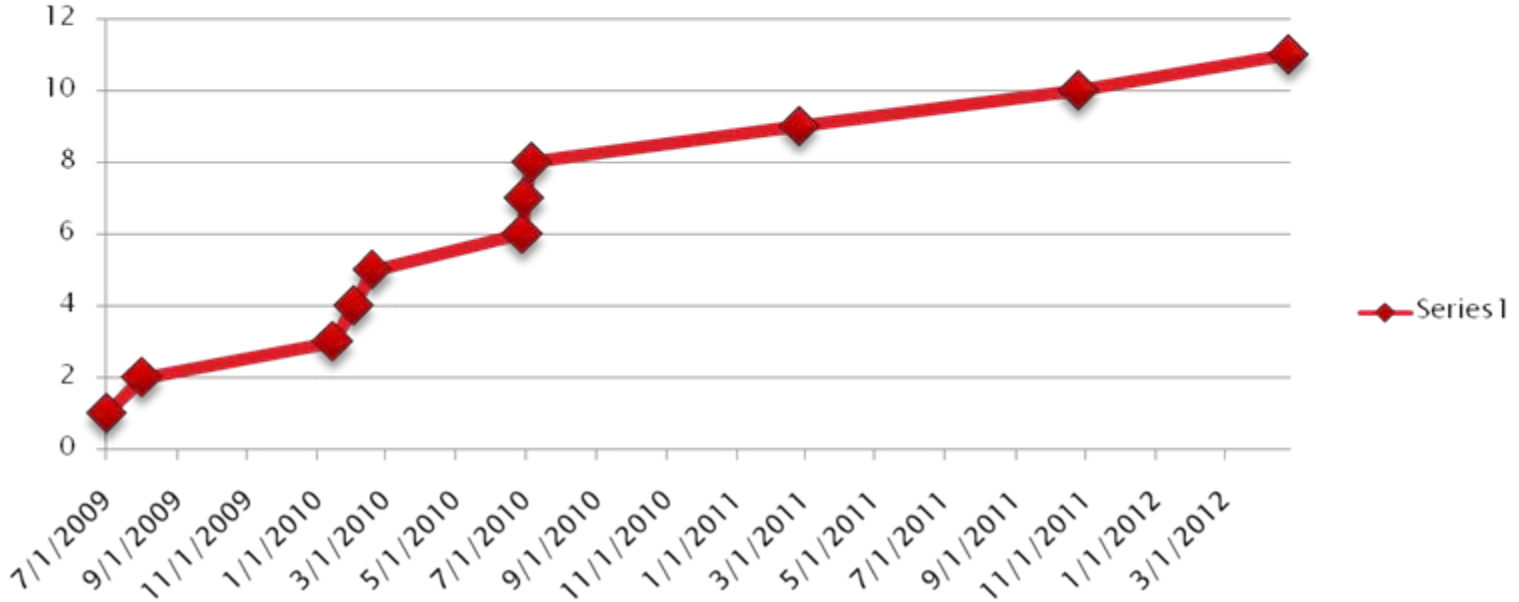
Conflict reports

- 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

Value normalization

- Different values for the same metadata
  - “inches” vs “2” vs “in.”
  - “Grayscale” vs “Greyscale”
- Different names for the same format
  - ‘JPEG2000’ vs ‘JPEG 2000’ vs ‘JPEG 2000 image”
- 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)

11 OS releases since July 2009



HARVARD LIBRARY

Code home

- http://fits.googlecode.com
  - Downloads: download the newest version
  - Mailing list: fits-users (new releases announced here)
  - Issues: File any bugs, upload patches

Future plans

- Support for container files & ContainerMD
  - arc.gz, .gz, .zip
- Improved video support
- Additional tools as needed
  - Apache Tika (docs, pdf, mbox, rtf, containers)
  - JHOVE 2 (shapefiles)
  - Mediainfo (audio, video)
  - Aduna Aperture (docs, pdf, email)
- Analysis of tool overlaps and “niches”
- Performance efficiencies
- Better documentation!