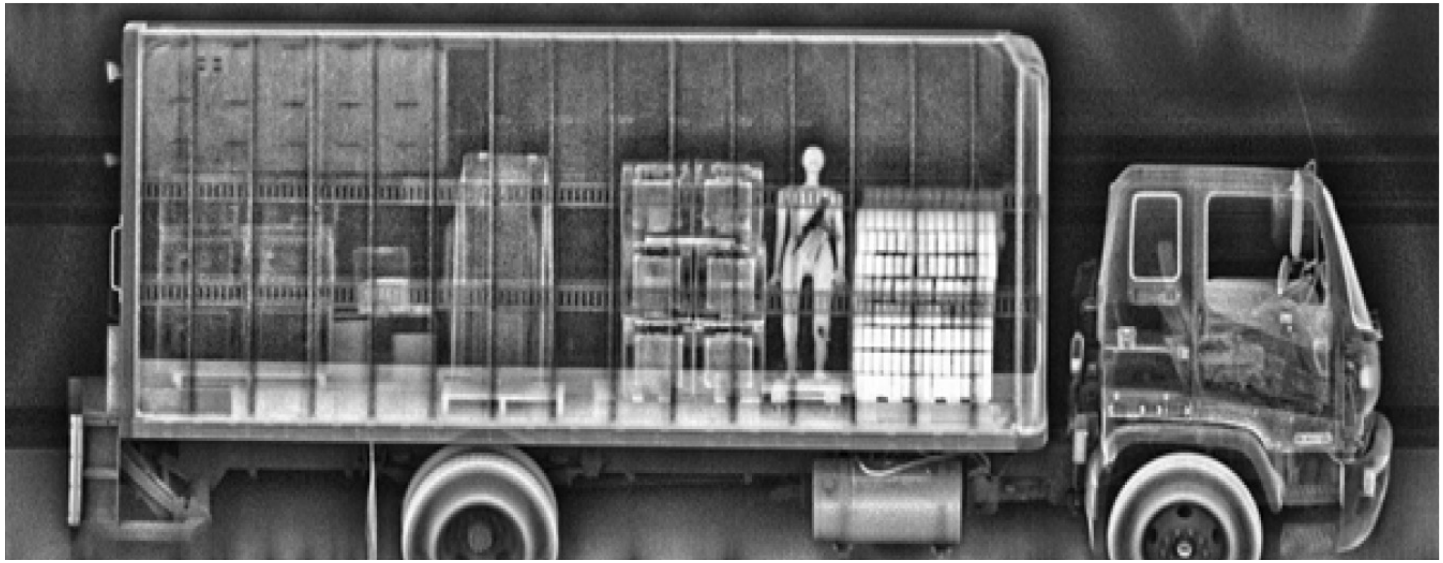
TLA-FLAT – FITS



KNAW Humanities Cluster /DI

Menzo Windhouwer

File Information Tool Set

- https://projects.iq.harvard.edu/fits
- "The File Information Tool Set (FITS) identifies, validates and extracts technical metadata for a wide range of file formats. 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."
- Tools like tika, droid, jhove, file, mediainfo
- Extensible (*not tried yet*)
- Configurable, e.g. to skip some
- CLI or servlet

FITS setup

- You can disable some tools you don't trust or are too expensive, e.g. recalculate a checksum
- Run it in a separate Tomcat with a restart policy, as mediainfo crashes from time to time

Courtesy Paul Trilsbeek

DoorKeeper action

- Run FITS on the resources to be ingested to
  - Confirm they are what they claim to be
  - Assert they have the right properties, e.g., bitrate or codec
  - Assign the correct MIME type
- Workflow
  - Call FITS to generate report XML
  - Find matching MIME types via XPaths
  - Check XPath assertions for each MIME Type
  - Accept MIME Type if all assertions are met

FITS report pt.1

KNAW Humanities Cluster /DI

- <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="1.4.0" timestamp="6/9/21 10:13 AM">
  <identification>
  <identity format="PDF/A" mimetype="application/pdf" toolname="FITS" toolversion="1.4.0">
  <tool toolname="Droid" toolversion="6.4"/>
  <tool toolname="Tika" toolversion="1.19.1"/>
  <version toolname="Droid" toolversion="6.4">2b</version>
  <externalIdentifier toolname="Droid" toolversion="6.4" type="puid">fmt/477</externalIdentifier>
  </identity>
  </identification>
  …

FITS report pt.2

KNAW Humanities Cluster /DI


```
…
<fileinfo>
      <size toolname="Jhove" toolversion="1.20.1">1528498</size>
      <creatingApplicationName toolname="Jhove" toolversion="1.20.1">Recoded by LuraDocument
PDF
         v2.53/Digitized by the Internet Archive</creatingApplicationName>
      <lastmodified toolname="Exiftool" toolversion="11.14" status="CONFLICT">2012:06:12
         20:53:46Z</lastmodified>
      <lastmodified toolname="Tika" toolversion="1.19.1" status="CONFLICT"
         >2012-06-12T20:53:46Z</lastmodified>
      <created toolname="Exiftool" toolversion="11.14" status="SINGLE_RESULT">2012:06:12
         20:53:12Z</created>
      <filepath toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT"
         >/app/flat/deposit/bags/test/bag-test-sip/data/test-sip/resources/Green Hornet
         003.pdf</filepath>
      <filename toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT">Green
         Hornet 003.pdf</filename>
      <md5checksum toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT"
         >c494e5d1850af32bbbc8b0d21d8d8a37</md5checksum>
      <fslastmodified toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT"
         >1598948540000</fslastmodified>
   </fileinfo>
   <filestatus/>
```


- …

FITS report pt.3

KNAW Humanities Cluster /DI

```
…
<metadata>
    <document>
      <subject toolname="Tika" toolversion="1.19.1" status="SINGLE_RESULT"
        >http://archive.org/details/GreenHornetComics3</subject>
      <title toolname="Tika" toolversion="1.19.1" status="SINGLE_RESULT">Green Hornet Comics #
        3</title>
      <author toolname="Tika" toolversion="1.19.1" status="SINGLE_RESULT">Fran
        Striker</author>
      <pageCount toolname="Tika" toolversion="1.19.1" status="SINGLE_RESULT">30</pageCount>
      <standard>
        <docmd:document xmlns:docmd="http://www.fcla.edu/docmd">
          <docmd:PageCount>30</docmd:PageCount>
        </docmd:document>
      </standard>
    </document>
  </metadata>
  <statistics fitsExecutionTime="2210">
    <tool toolname="MediaInfo" toolversion="0.7.75" status="did not run"/>
    <tool toolname="OIS Audio Information" toolversion="0.1" status="did not run"/>
    <tool toolname="ADL Tool" toolversion="0.1" status="did not run"/>
    <tool toolname="VTT Tool" toolversion="0.1" status="did not run"/>
    <tool toolname="Droid" toolversion="6.4" executionTime="696"/>
    <tool toolname="Jhove" toolversion="1.20.1" executionTime="2158"/>
    <tool toolname="file utility" toolversion="5.38" executionTime="2172"/>
    <tool toolname="Exiftool" toolversion="11.14" executionTime="2070"/>
    <tool toolname="NLNZ Metadata Extractor" toolversion="3.6GA" executionTime="2048"/>
    <tool toolname="OIS File Information" toolversion="0.2" executionTime="679"/>
    <tool toolname="OIS XML Metadata" toolversion="0.2" status="did not run"/>
    <tool toolname="ffident" toolversion="0.2" executionTime="2054"/>
    <tool toolname="Tika" toolversion="1.19.1" executionTime="1748"/>
  </statistics>
</fits>
```

Example 1

```
<mimetype value="audio/x-wav">
    <assertions xpath="//fits:identity[fits:tool/@toolname = 'OIS Audio Information']/@format = 'Waveform Audio'">
        <assert xpath="/fits:fits/fits:fileinfo/fits:filename/replace(., '.*\.', '') = 'wav'" message="Your file [{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] is a WAV file, but has a different file extension than '.wav'"/>
        <assert xpath="/fits:fits/fits:metadata/fits:audio/fits:sampleRate = ('44100', '48000')" message="The sample rate of your audio file [{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] is [{/fits:fits/fits:metadata/fits:audio/fits:sampleRate}], should be 44100 or 48000"/>
        <assert xpath="/fits:fits/fits:metadata/fits:audio/fits:bitDepth = ('16', '24')" message="The bit depth of your audio file [{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] is [{/fits:fits/fits:metadata/fits:audio/fits:bitDepth}], should be 16 or 24"/>
        <assert xpath="/fits:fits/fits:metadata/fits:audio/fits:channels = ('1', '2')" message="Your audio file [{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] has [{/fits:fits/fits:metadata/fits:audio/fits:channels}] channels, should be 1 or 2"/>
        <assert xpath="/fits:fits/fits:metadata/fits:audio/fits:audioDataEncoding[@toolname = 'OIS Audio Information'] = 'PCM'" message="Your audio file [{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] has the encoding [{/fits:fits/fits:metadata/fits:audio/fits:audioDataEncoding[@toolname = 'OIS Audio Information']}], should be 'PCM'"/>
    </assertions>
</mimetype>
```

Courtesy Paul Trilsbeek

Example 2

```
<mimetype value="text/x-eaf+xml"
xpath="//fits:identity/@mimetype='text/xml'">
        <assertions xpath="/fits:fits/fits:fileinfo/fits:filename/replace(., '.*\.', '') =
'eaf'">
            <assert
xpath="exists(/fits:fits/fits:metadata/fits:text/fits:markupLanguage[starts-
with(., 'http://www.mpi.nl/tools/elan/EAFv')])" message="Your ELAN file
[{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] does not contain a valid
schema declaration"/>
            <assert xpath="/fits:fits/fits:metadata/fits:text/fits:charset = 'UTF-8'"
message="The character encoding of your ELAN file
[{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] is
[{/fits:fits/fits:metadata/fits:text/fits:charset}], should be UTF-8"/>
    <assert xpath="/fits:fits/fits:filestatus/fits:valid = 'true'" message="Your
ELAN file [{replace(/fits:fits/fits:fileinfo/fits:filename,$work,'')}] is not a valid
XML file"/>
        </assertions>
</mimetype>
```

Courtesy Paul Trilsbeek

To share or not to share?

KNAW Humanities Cluster /DI

- Assertions to determine proper MIME type for too generic MIME types, e.g. application/xml and video/mpeg
- Java lib/cli to validate assertions
- FITS extensions, if any

KNAW Humanities Cluster /DI

Questions?

menzo.windhouwer@di.huc.knaw.nl

@mwindhouwer