MUPPET: Multi Pass File Properties Extraction Tool

Pitch Document CONCEPT

Maurice de Rooij, Bill Roberts, Mette van Essen and Maurice van den Dobbelsteen, Nationaal Archief - programme Digital Preservation - October 2011

INTENDED AUDIENCE:

Content Owners and Engineers.

THE PROBLEM:

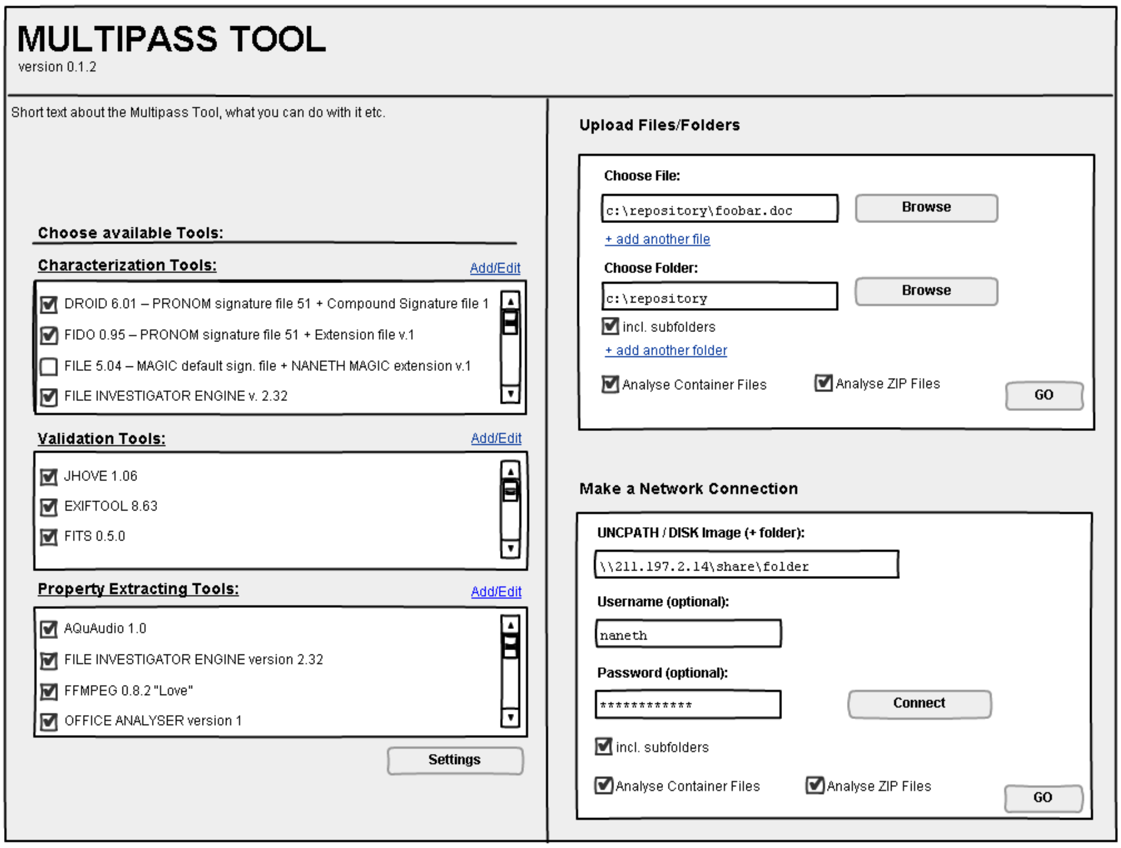
Currently there is not a single tool which can identify files and extract all possible properties at the same time. Some tools come close, but have their limitations. One solution could be to pass a file (or files) by hand through multiple tools but this is a tedious job. Next to that markup and the amount of output is very different: sometimes too brief, sometimes too verbose.

THE APPROACH:

Create a lightweight/simple API, which incorporates necessary tools by deploying a simple wrapping mechanism. Some of the tools for the first pass could include: FIDO, DROID and File Investigator Engine (commercially licensed) to determine what format is it. This is necessary for the second pass: to have the files analyzed by property extractor tools like ImageMagick identify (image characterization), AQuAdio (audio characterization) and Office Analyser (Word document characterization). Output from all tools should be normalized to a generic grammar in such a way that properties can be evaluated against each other. On top of this API there should be a simple web based GUI to invoke the tools. The GUI should merely provide a way to make the tool more user friendly. This also implies that the API can also be used from the command line, generating the same results. Results are to be saved in a database from which reports can be generated in HTML, PDF and XML/RDF or can be used in any other way. Main purpose is to provide content owners with a tool to perform a preliminary “quick scan” of content, as well to provide engineers a way to have their repositories scanned for more in-depth analysis.

CHALLENGES:

1. Create a lightweight/simple API to command tools
2. Create a simple wrapping mechanism to add tools in a simple way
3. Create a generic grammar to save results, using Planets Ontology as basis
4. Create a translator to map tool output to generic grammar
5. Create a user friendly GUI on top of the API without bells and whistles



MULTIPASS TOOL

version 0.1.2

Short text about the Multipass Tool, what you can do with it etc.

Choose available Tools:

Characterization Tools:

☑ DROID 6.01 – PRONOM signature file 51 + Compound Signature file 1

☑ FIDO 0.95 – PRONOM signature file 51 + Extension file v.1

□ FILE 5.04 – MAGIC default sign. file + NANETH MAGIC extension v.1

☑ FILE INVESTIGATOR ENGINE v. 2.32

Validation Tools:

☑ JHOVE 1.06

☑ EXIFTOOL 8.63

☑ FITS 0.5.0

Property Extracting Tools:

☑ AQuAudio 1.0

☑ FILE INVESTIGATOR ENGINE version 2.32

☑ FFMPEG 0.8.2 "Love"

☑ OFFICE ANALYSER version 1

Upload Files/Folders

Choose File: `c:\repository\foobar.doc`

Choose Folder: `c:\repository`

☑ incl. subfolders

☑ Analyse Container Files ☑ Analyse ZIP Files

Make a Network Connection

UNCPATH / DISK Image (+ folder): `\\211.197.2.14\share\folder`

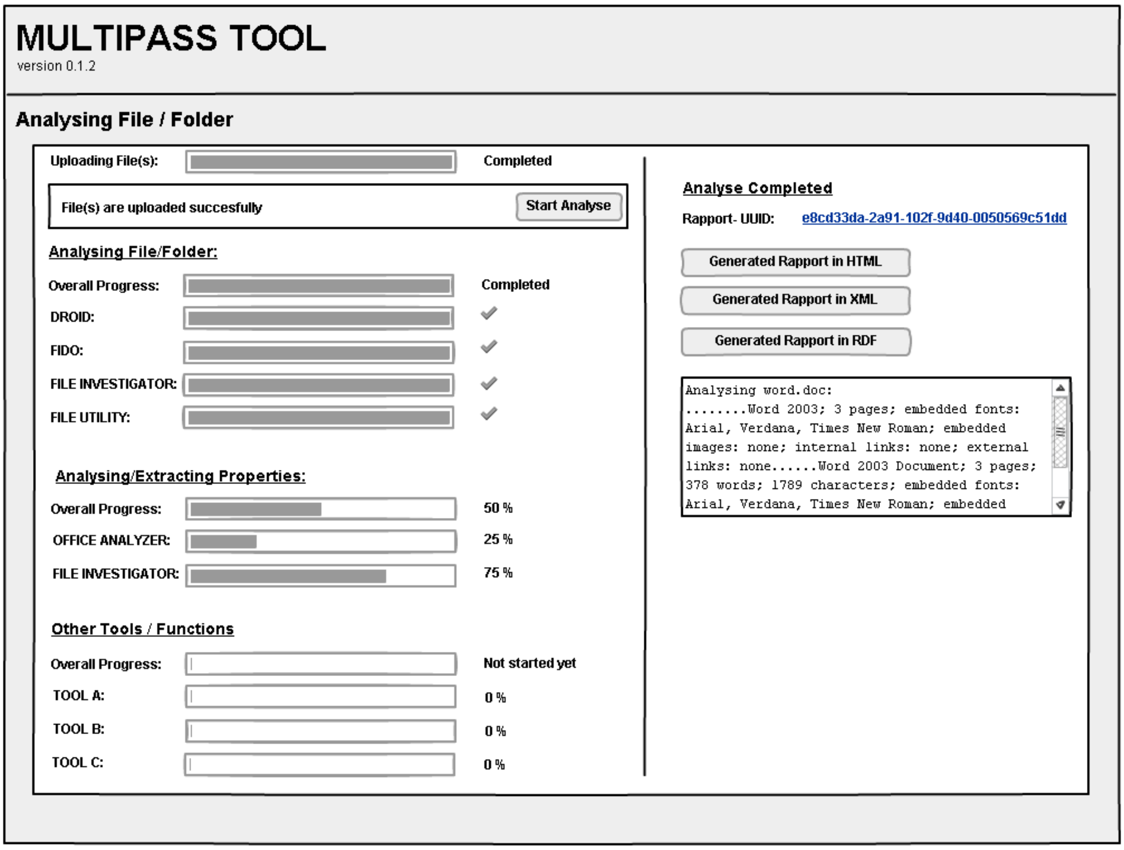
Username (optional): `naneth`

Password (optional): `************`

☑ incl. subfolders

☑ Analyse Container Files ☑ Analyse ZIP Files

Main screen of MUPPET



MULTIPASS TOOL

version 0.1.2

Analysing File / Folder

Uploading File(s): Completed

File(s) are uploaded succesfully

Analysing File/Folder:

Overall Progress: Completed

DROID: ✓

FIDO: ✓

FILE INVESTIGATOR: ✓

FILE UTILITY: ✓

Analysing/Extracting Properties:

Overall Progress: 50 %

OFFICE ANALYZER: 25 %

FILE INVESTIGATOR: 75 %

Other Tools / Functions

Overall Progress: Not started yet

TOOL A: 0 %

TOOL B: 0 %

TOOL C: 0 %

Analyse Completed

Rapport- UUID: e8cd33da-2a91-102f-9d40-0050569c51dd

```
Analysing word.doc:
........Word 2003; 3 pages; embedded fonts:
Arial, Verdana, Times New Roman; embedded
images: none; internal links: none; external
links: none......Word 2003 Document; 3 pages;
378 words; 1789 characters; embedded fonts:
Arial, Verdana, Times New Roman; embedded
```

Example screen during analysis