

Tool Demo: Making Sense of a Collection



Digital**Preservation**Coalition

File formats are a problem for preservation ... sort of



Changing file formats
vs
'Robust' formats
vs
Proliferating formats
vs
Conformant data containers

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

What's in a File?



```
10101010101111010001010101000
100101001011110100101010101001
00100101010100100000101010101
01111101000101010100010010100
10111010010101010100100100101
01010010000010101010101111101
00010101010001001010010111010
01010101010010010010101010010
00001010101010110101010111101
00010101010001001010010111010
01010101010010010010101010010
00001010101010111110100010101
01000100101001011101001010101
01001001001010101001000001010
1010101111010001010101000100
1010010111010010101010100100...
```

```
SOI
APP0 JFIF
1.2
APP13 IPTC
APP2 ICC
DQT
SOF0 200x392
DRI
DHT
SOS
ECS0
RST0
ECS1
RST1
ECS2...
```



Scale = automation = software tools



- Digital typically means working at scale
- Automation becomes a necessity
- Software tools
- Techie stuff
- But...
- Hands on experience



Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Making Sense of a Collection



- Understand the data, then assess risks, plan, take action to preserve
- Characterisation:
 - How many files?
 - How big are the files?
 - What file formats?
 - Is the data dynamic or interactive?
 - Does it contain personal information?
 - Is it encrypted?



Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

File Formats: What Flavour is Your Data?



- What are they likely to contain?
- What are the likely preservation risks associated with that file format?
- What software do I need to render, interpret or make sense of that file format?
- Caveat: Fuzzy concept, doesn't give all the answers
 - [Abusing file formats](#) by Ange Albertini (chapter 6)

PRONOM and DROID

Pronom: a register of file formats and their behaviours (probably the world's most boring database)

DROID: a tool that analyses the files on a system (using the most boring database in the world)



Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

PRONOM

<http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>

DROID

<https://sourceforge.net/projects/droid/>

File Formats and Their Characteristics



The National Archives

 [Advanced search](#)

[About us](#) | [Education](#) | [Records](#) | [Information management](#) | [Shop online](#)

You are here: [Home](#) > [Services for professionals](#) > [Preservation](#) > PRONOM



The technical registry
PRONOM

Welcome | About | Add an entry
Search | Help | Information resources

Welcome to PRONOM

[PRONOM changes and DROID signature file release notes](#)

Find out more about our plans to make PRONOM's data available online. PRONOM is a register of technical information, PRONOM is a register of software products and other technical components required for historical or business value. Find out about the future of PRONOM.

[Simple search](#) | [File format](#) | [PRONOM Unique Identifier](#) | [Software](#) | [Vendor](#) | [Lifecycle](#) | [Migration Pathways](#)

Details for: **Tagged Image File Format 3**

[Save as...](#) | [XML](#) | [CSV](#) | [Print](#)

Go to: [Summary](#) | [Documentation](#) > | [Signatures](#) > | [Compression](#) > | [Character encoding](#) > | [Rights](#) > | [Reference files](#) > | [Properties](#) >

Summary

| | |
|-----------------------|--|
| Name | Tagged Image File Format |
| Version | 3 |
| Other names | TIFF (3) |
| Identifiers | PLUID: fmf/7 MIME: image/tiff Apple Uniform Type Identifier: public.tiff |
| Family | Image (Raster) |
| Classification | Image (Raster) |
| Disclosure | Full |
| Description | The Tagged Image File Format (TIFF) is a raster image format originally developed by the Aldus Corporation, primarily for use in scanning and desk-top publishing. When Adobe Systems Incorporated purchased Aldus in 1994, they acquired the rights to the TIFF format and have maintained it since then. TIFF files comprise three sections: an Image File Header (IFH), an Image File Directory (IFD), and the image data. TIFF files can contain multiple images (multi-page TIFF), and each image has a separate IFD. The IFH always appears at the beginning of the file, and is immediately followed by a pointer to the first IFD. The IFD contains metadata which describes the associated image, stored as a series of tags. The IFD also contains a pointer to the actual image data. TIFF 3.0 supports colour depths from 1 bit to 24 bit (e.g. monochrome to true colour), and a range of compression types (LZW, CCITT Group 3 and Group 4). |



DROID: Search and report on files from an entire network

Identify files by 'extension'

Identify files by 'signature'

Identify subtypes and versions

Report errors and concerns

Scalability ...

Built for TNA ...

Only as good as data ...



Also in this space:

- C3PO
- JHOVE
- TIKKA
- FITS



Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

*Only as good as
their data:
SO JOIN IN!*

- CRISP
- JUST FIX IT
- PRONOM
- ...

JHOVE - <http://jhove.sourceforge.net/>

C3PO - <http://www.scape-project.eu/tools>

FITS - <http://fitstool.org>



TRUST NO ONE

Assume nothing, validate everything



Stuff happens

- Whenever a digital collection is moved, processed, curated or altered in any way.... things can go wrong!
 - Network dropouts at critical times
 - Disks get full, subsequent data copied there is lost
 - Software bugs lead to unexpected results
 - Human error leads to all sorts of issues

- Stuff happens a lot more at scale!