Tool Demo: Making Sense of a Collection



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?



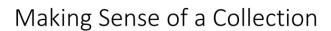
SOI
APPO JFIF
1.2
APP13 IPTC
APP2 ICC
DQT
SOFO 200x392
DRI
DHT
SOS
ECSO
RSTO

ECS1

RST1

ECS2...







- 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?
 - What risks are associated?
- Scale = automation = software tools



Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Digital typically means working at scale Automation becomes a necessity 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)

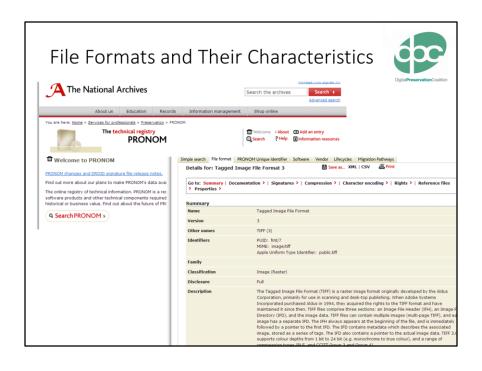


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

PRONOM

http://www.nationalarchives.gov.uk/PRONOM/Default.aspx DROID

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







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



Only as good as their data: SO JOIN IN!

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

Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

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

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

FITS - http://fitstool.org





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!