DigitalPreservationCoalition

Tool demo: making sense of a collection

Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

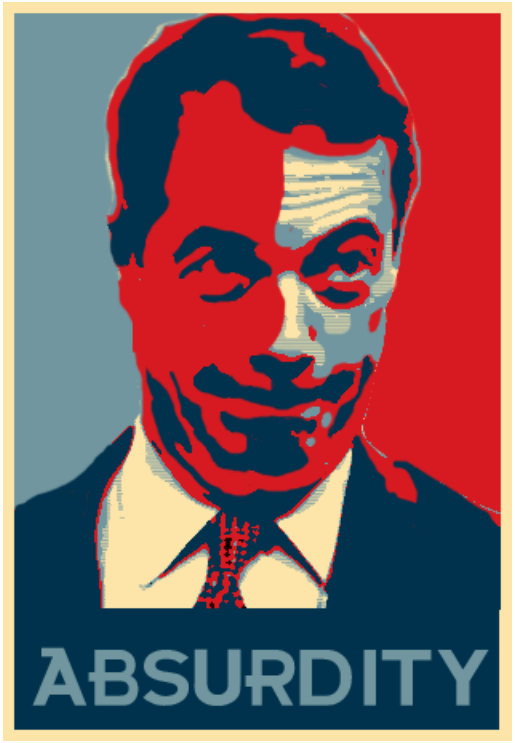
Scale = automation = software tools

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

Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

DigitalPreservationCoalition

10101010101111010001010101000	SOI
10010100101111010010101010101001	APP0 JFIF 1.2
00100101010100100000101010101	APP13 IPTC
01111101000101010100010010100	APP2 ICC
10111010010101010100100100101	DQT
01010010000010101010101111101	SOF0 200x392
00010101010001001010010111010	DRI
01010101010010010010101010010	DHT
00001010101010110101010111101	SOS
00010101010001001010010111010	ECS0
01010101010010010010101010010	RST0
00001010101010111110100010101	ECS1
01000100101001011101001010101	RST1
01001001001010101001000001010	ECS2…
10101011111010001010101000100	
10100101110100101010100100…	



Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

DigitalPreservationCoalition

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?

Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

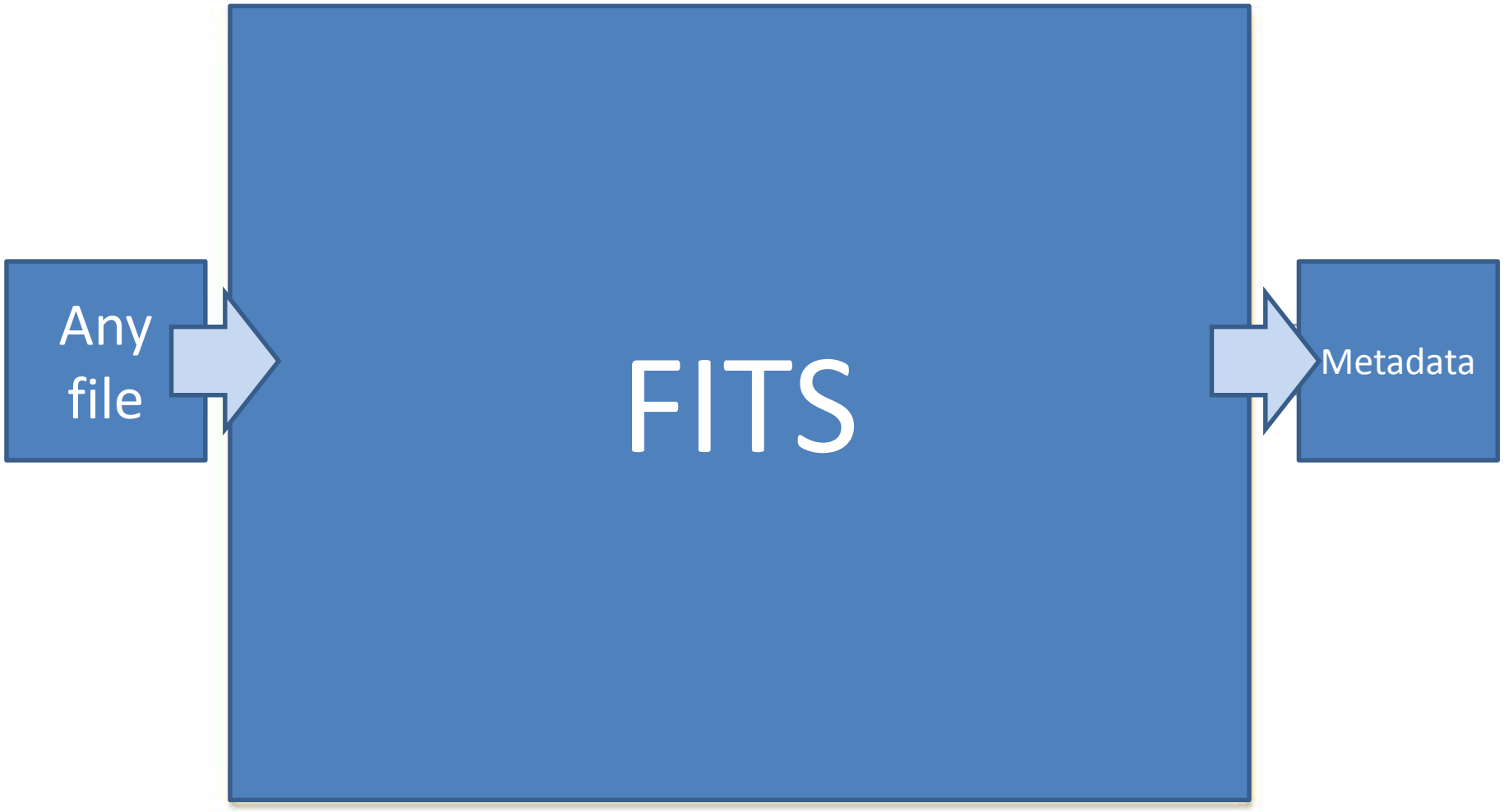
DigitalPreservationCoalition

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)

Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

FITS – File Information Tool Set



Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

DigitalPreservationCoalition

C3PO

- FITS: Content profiling tool
- Works with FITS to visualize FITS output metadata
- Setup can be a little more taxing

Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley

Assume nothing, validate everything

DigitalPreservationCoalition

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!

Paul Wheatley
paul.wheatley@dpconline.org
@prwheatley