ads Archaeology Data Service

Migration and Normalisation

TNA Training School

13th November 2024

Olivia Foster

ads Archaeology Data Service

Migration and Normalisation

- What is migration and normalisation?
- Creating a migration and normalisation strategy
- Migration and normalisation at the Archaeology Data Service

https://doi.org/10.5281/zenodo.11147887

ads Archaeology Data Service

File Formats

- Different formats for different data types (e.g. images, audio, 3D data)
- Preservation implications for different formats:
    - Hardware or Software dependencies
    - Open source vs proprietary
    - Ubiquity

https://digitalbevaring.dk/

Digital Preservation Strategies

One approach to digital preservation is to migrate data from one format to another to preserve **content**.

Other approaches include:

- Replication
- Refreshing
- Emulation (re-creating the original operating environment to preserve look and feel of a resource)

Migration

n.

The process of moving data from one information system or storage medium to another to ensure continued access to the information as the system or medium becomes obsolete or degrades over time.

– Dictionary of Archives Terminology

ads Archaeology Data Service

Why Migrate Files?

XLS
MIGRATION
CSV

- Can include moving between formats or physical media
- Migration due to concerns over **obsolescence**
- **Preserves content** to maintain access
- To make data more **accessible** (e.g. open source software)

However It is likely that the majority of file formats you deal with will be commonly understood and well supported

Normalisation

n.

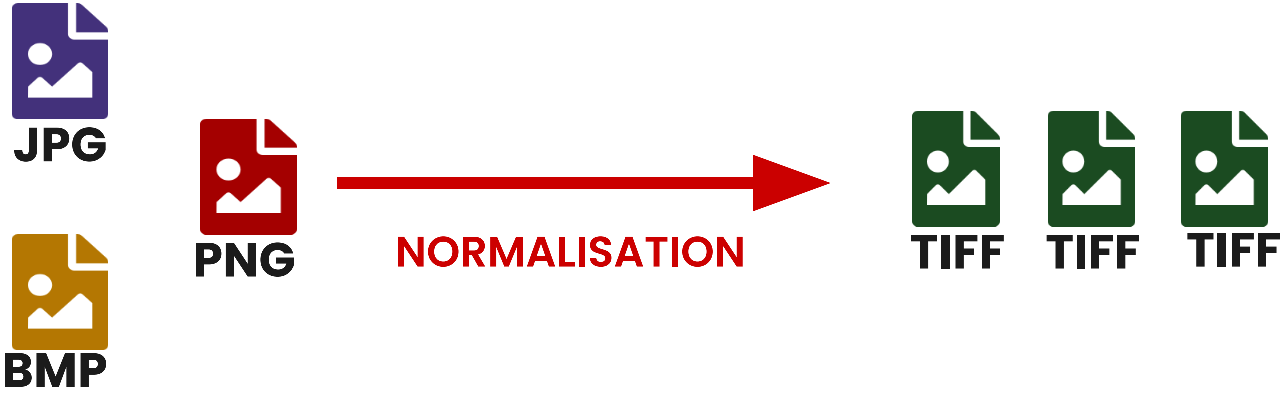
the process of converting a digital object into a persistent file format

– Dictionary of Archives Terminology

Archaeology Data Service

What is Normalisation?

- Migrating to a standardized format (e.g images to uncompressed TIFF)
- A Persistent file format is selected to preserve data because it is expected to remain usable, reliable, and accessible over a long period of time



ads Archaeology Data Service

Why Normalise Files?

- Fewer formats = less complexity
- Can be many versions of some formats (e.g. JPG)
- Normalised format needs to be selected carefully!

Extension	File Type	File Type Version	PRONOM ID
JPG	JPEG File Interchange Format	1.02	fmt/44
JPG	Raw JPEG Stream		fmt/41
JPG	Exchangeable Image File Format (Compressed)	2.2	x-fmt/391
JPG	Exchangeable Image File Format (Uncompressed)	2.2	x-fmt/387
JPG	Exchangeable Image File Format (Compressed)	2.3 x	fmt/1507
JPG	JPEG File Interchange Format	1.00	fmt/42
JPG	Nikon Digital SLR Camera Raw Image File		fmt/202

To simplify …

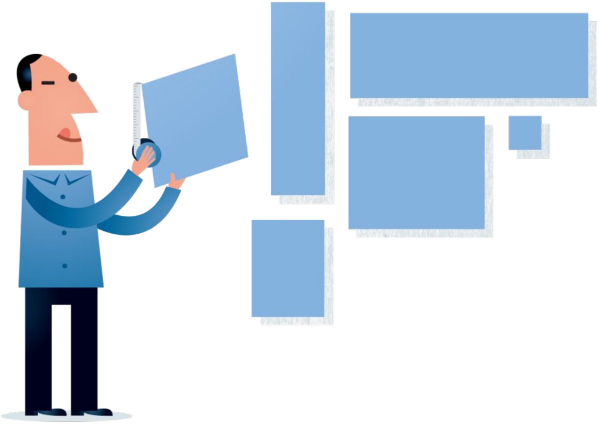
Migration to avoid file format obsolescence

Normalisation to avoid file format proliferation

Selecting Formats

Formats selected should best meet the requirements of the collection content and preserves the qualities of the content. A few things to consider:

- Open source vs proprietary
- Ubiquity (how widely used)
- Compression vs uncompressed
- Documentation and standards
- Different needs for preservation and access
- What are other similar organisations doing?



Digitalbevaring.dk

Archaeology Data Service

Resources for Selecting Formats

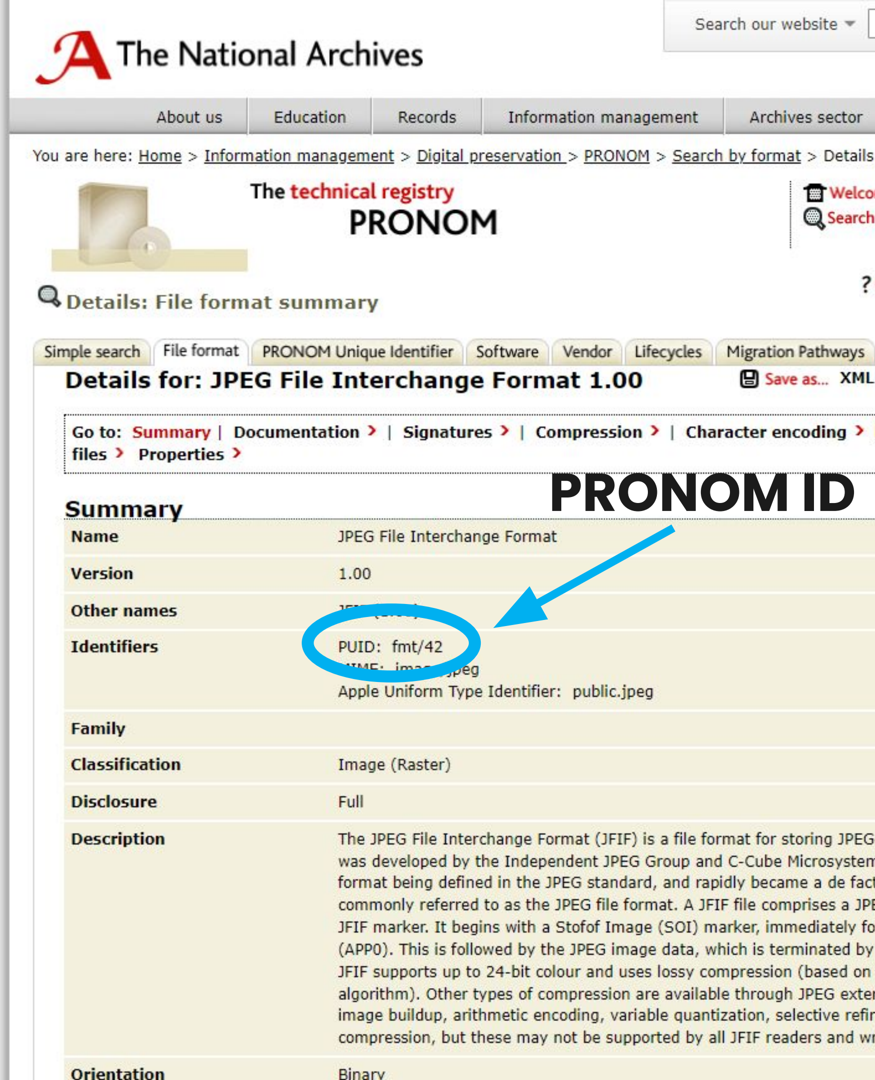
- DPC’s 'Bit List' of Endangered Digital Species
- Library of Congress recommended format specifications
- OPF File Format Risk Registry
- PRONOM

https://www.dpconline.org/digipres/champion-digital-preservation/bit-list

ads Archaeology Data Service

Carrying out Migrations

- Manual or automated processes (e.g. XnView)
- Validation (to check migration was successful and data hasn't changed)
- Document actions taken (and why)
- Tools to identify formats:
  - DROID
  - PRONOM



You are here: Home > Information management > Digital preservation > PRONOM > Search by format > Details

The technical registry PRONOM

Details: File format summary

Details for: JPEG File Interchange Format 1.00

Go to: Summary | Documentation › | Signatures › | Compression › | Character encoding › files › Properties ›

Summary

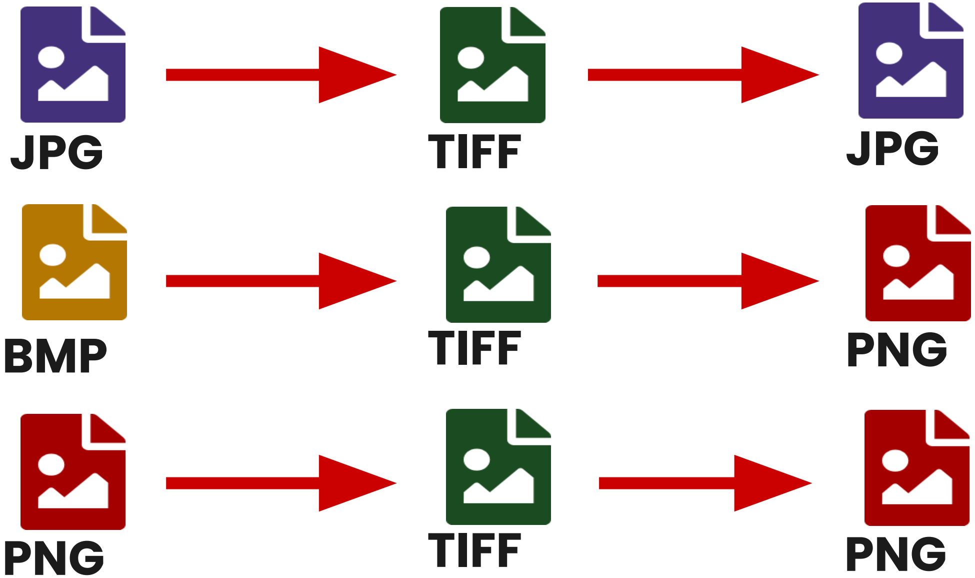
Name	JPEG File Interchange Format
Version	1.00
Other names	JFIF (1.00)
Identifiers	PUID: fmt/42 MIME: image/jpeg Apple Uniform Type Identifier: public.jpeg
Family	
Classification	Image (Raster)
Disclosure	Full
Description	The JPEG File Interchange Format (JFIF) is a file format for storing JPEG was developed by the Independent JPEG Group and C-Cube Microsystem format being defined in the JPEG standard, and rapidly became a de fact commonly referred to as the JPEG file format. A JFIF file comprises a JPE JFIF marker. It begins with a Stofof Image (SOI) marker, immediately fo (APP0). This is followed by the JPEG image data, which is terminated by JFIF supports up to 24-bit colour and uses lossy compression (based on algorithm). Other types of compression are available through JPEG exter image buildup, arithmetic encoding, variable quantization, selective refin compression, but these may not be supported by all JFIF readers and wr
Orientation	Binary

Creating a Migration and Normalisation Strategy

- Identifying formats commonly used for content (or already in repository)
- Assessing preservation risks to formats
- Identify preferred formats for preservation (and also access and ingest)
- Carry out migrations and/or normalisation
- Record these preservation actions
- Review formats in repository periodically

Migration and normalisation at the ADS: Images

- Original files are deposited in range of formats
- Preservation versions of files created by migrating to a preferred **preservation format** (normalisation)
- Dissemination versions of files created by either **replicating** original files or migrating to preferred **dissemination format**



JPG → TIFF → JPG

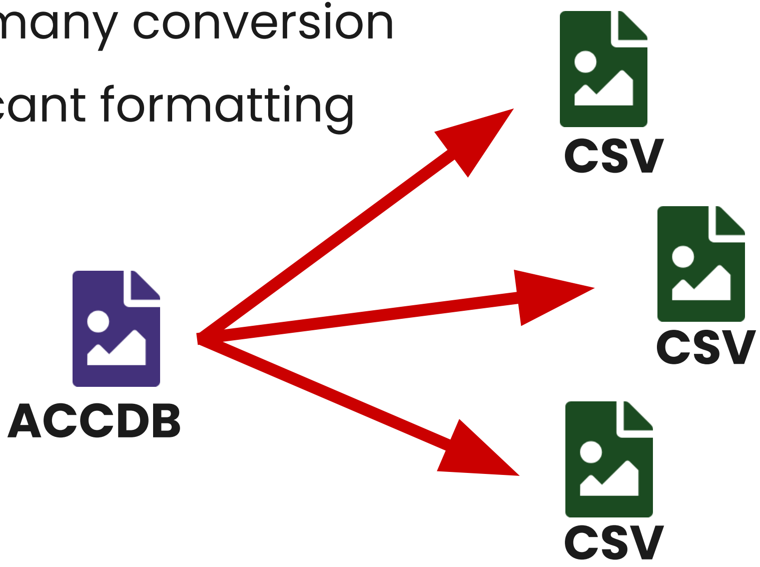
BMP → TIFF → PNG

PNG → TIFF → PNG

Archaeology Data Service

Migration and normalisation at the ADS: Databases

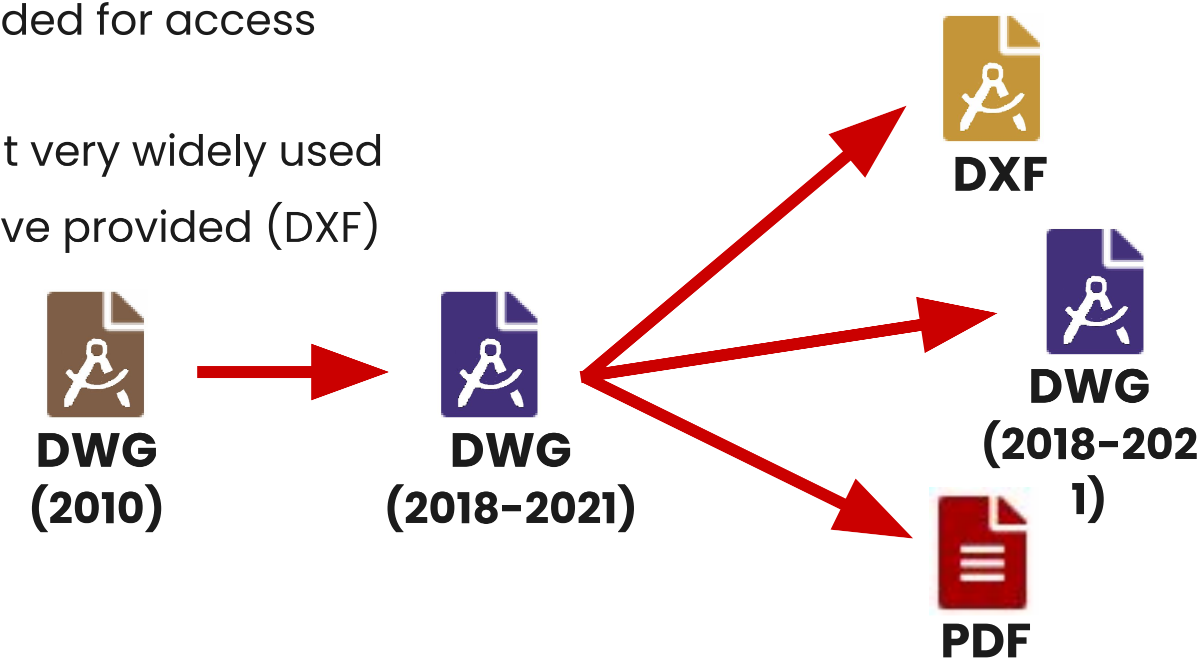
- Migration may involve one to many conversion
- Consider data loss, e.g. significant formatting



ads Archaeology Data Service

Migration and normalisation at the ADS: CAD

- Multiple formats provided for access (dissemination)
- DWG is proprietary but very widely used
- Open source alternative provided (DXF)



Resources:

- DPC Digital Preservation handbook: File formats and standards https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards
- The Global 'Bit List' of Endangered Digital Species - https://www.dpconline.org/digipres/champion-digital-preservation/bit-list
- PRONOM Technical Registry - https://www.nationalarchives.gov.uk/PRONOM/
- Library of Congress: Sustainability of Digital Formats - https://www.loc.gov/preservation/digital/formats/
- DROID (Digital Record Object Identification) tool developed by The National Archives - https://www.nationalarchives.gov.uk/information-management/manage-information/policy-process/digital-continuity/file-profiling-tool-droid/

ads Archaeology Data Service

Any questions?

Archaeology Data Service
Department of Archaeology
University of York
The King's Manor
Exhibition Square
York, YO1 7EP

www.archaeologydataservice.ac.uk

help@archaeologydataservice.ac.uk