ICA Africa Programme
Programme pour l'Afrique

ICA Training Programme
Programme de Formation

Digital Records Curation Programme

Week 8:
Planning for Digitisation

Week 7 Recap

What did you learn?

- Class on Digital Repositories and Digital Records Appraisal
- Tutorials on Using Digital Preservation Software

Learning outcomes:

At the end of this class you will be able to:

- explain the organisational, logistical, legal and technical issues involved in digitisation
- understand the basics of image file formats, particularly resolution and losslessness
- understand what optical character recognition is

Digitisation

- Digitisation is the conversion from analogue to digital formats
- Digitisation is typically done by scanning or photography
- Digitisation creates ‘digital surrogates’ not ‘digital records’

Why digitise?

The National Archives (UK) advises that digitisation can support:

- Access
- Generating income
- Brand
- Searchability
- Preservation
- Interaction
- Integration
- Disaster recovery

Digitisation decision-making matrix

	Question 1	Question 2	Question 3	Question 4	Question 5
Assessment	Is there user support?	What are the local collection development policies?	Does this form a national or international contribution?	Does a similar product already exist elsewhere?	Is this conservation or preservation?
Gains	Does digitisation reduce wear on the originals or open up access?	Is the intellectual content of the work enhanced?	Is navigation easy?	Are disparate collections unified?	Is use of the damaged original material enriched?
Standards	Have suitable standards been followed?	Are the originals available from a variety of hardware platforms?	Is the software available and easy to use?	Does the metadata conform to agreed standards?	What are the archiving requirements?
Administrative Issues	Do you have enough money?	Have copyright and rights issues been secured?	Does your institution have enough expertise?	Is there a partnership with a commercial provider?	Do the benefits justify the costs?

You should be familiar with…

- The legal issues
- Metadata considerations
- Preservation of digital surrogates

What about:

- Retention and disposition of originals?
- Retention and disposition of surrogates?
- Workflow and quality control?
- Ongoing costs?

Group Work: Digitisation Decision Tree

- Consider an organisation you have worked for. Is the organisation ready for digitisation?
- What policy frameworks need to be in place?
- What technical infrastructure would need to be in place before they proceed?

Image File Formats

There are two basic types of image format:

- vector graphics, which contain resizable shapes and lines without loss of edge definition (and the more advanced also support 3-dimensional rotation, etc)

- raster images (or bitmaps), which are pictures formed of minute pixels of tone, from monotone (B&W), through greyscale and colour to true (24-bit) colour.

Extension	Name	Variants	Characteristics
.tiff	Tagged image fixed format	Sophisticated format with a very large number of different options	Adobe-owned format but specification stable since 1992 and with a wide variety of supporting applications. Some sub-types use lossless compression. Can support very large, professional /commercial quality images. **Best option for long-term preservation if used correctly.**
.jpeg	Joint photo-graphers' expert group	JPEG 2000 is the latest variant	Lossy Compression format produced by most handheld digital cameras; compresses more each time a file is saved. **Web friendly but not suitable for long term use**
.pdf	Portable document format	PDF-A [Archival] is a variant of PDF1.4 and the specification is ISO 19005	ADOBE Corp. - owned proprietary formats, but ADOBE's policy is to make previous version specifications available after new releases,Default save format of Acrobat software and lots of free viewers available, but likely that PDF-A will spawn other creating software too. Compliant with Postscript protocols [ADOBE-owned but open]
.png	Portable network graphics	ISO 15948: 2003	Web friendly format developed as an open standard to avoid intellectual property disputes in use of Compuserve-owned *.gif* format; Mainly designed for web use Uses lossless compression

Resolution and Loss

- Image resolution
- Lossy versus lossless
- Preservation copies versus reference copies
- Storage costs go up with file sizes

Optical Character Recognition

OCR is a process that recognises the shape of standard fonts and converts an image of a page of analogue text to an electronic file of usable text. If scanned images are to be searchable or the text used in other ways there are consequences for how the images are stored, indexed and retrieved in the future, namely:

- Some image files recognise textual information as well as treating text as a picture (e.g. ADOBE *.pdf*)
- Other image formats will not provide this facility. This leaves a choice of whether to save a text file – produced using OCR - in close association with the scanned image to provide whole-content searchability or to index the content in some other way (e.g. through metadata).
- OCR requires quality control measures at the scanning stage for future search results to be reliable. OCR technology can claim accuracy rates that sound impressively high in percentage terms, but a 2% error rate is a lot of errors in a big project!

Digitisation in Practice

- Insert the guest lecturer’s slides here

Conclusion

- Digitisation: costs, logistics, legal issues, technical requirements and choices
- Selecting image file formats for digitisation ‘masters’ and reference copies
- OCR

Any questions?