Quality Assured Image Migration

Scalable Workflow Background

Peter May
Digital Preservation Technical Architect, British Library

Future Formats First, Building Applications Infrastructure for Action Services
British Library, 16th September 2013
Outline

• Setting the scene: why it’s important to look at image migration workflows?
  • Collection: Digitised Newspapers
  • Collection concerns

• Outline conceptual workflow
Collection: British Newspapers

• Large collection of British newspapers
  • Old newspapers are fragile
  • Digitised through various JISC funded projects for preservation and access reasons
  • >2million digitised images (TIFFs)
  • 1620-1900’s

• The British Newspaper Archive
  • Partnership with brightsolid
  • ~7 million digitised scans in collection (incl. from JISC projects)
  • Up to 8000 scans per day
  • Files as large as 400MB
  • http://www.britishnewspaperarchive.co.uk
Collection Concerns

• Need to store this collection (into the future)
  • Cost effectively
    • Large files -> more storage -> greater costs
    • Smaller files -> less storage required -> cost savings
    • Cost savings potentially leads to money to invest in...
      • ... digitising more
      • ... buying more storage
      • ... refreshing digitisation equipment

• Need to provide efficient access to this collection also
  • At varying levels (navigation->detail)
  • Without a large increase in storage expense
    • multiple resolution images increase storage costs

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Digital Access Use Cases

a) **Navigation**: display of thumbnail images from multiple master files

b) **Reading**: display at an intermediate “reading” resolution a single master with zoom and pan (and occasionally two pages side-by-side)

c) **Detailed**: display at full resolution with pan

- **Observation**: (a) and (b) will be much more frequent than (c)
Access Use Cases

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Solution: Migrate the files?

- Initial investigations indicated that migrating TIFFs to JPEG2000 could:
  - Reduce storage size/costs
  - Facilitate enhanced user access

- However, cost savings can only be realised through deleting original TIFFs
  - Must ensure the quality and validity of the migration
    - Avoiding (or at least detecting) corrupted migrated images

- Secondly, we have a lot of files to process
  - How to do this in an efficient and scalable way?

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Conceptual Migration Workflow - Tools

- Migration Codecs:
  - Kakadu
  - OpenJPEG
  - ...

- QA/Validation tools:
  - Jpylyzer
  - Matchbox (image feature analysis)
  - Exiftool (metadata extraction)
Conceptual Migration Workflow - Validation

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Conceptual Migration Workflow - QA

Extract from originals:
- Technical Metadata (ExifTool)
- Image Features (Matchbox)

Extract from migrated:
- Technical Metadata (ExifTool)
- Image Features (Matchbox)

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Concluding remarks

• Important to investigate migration workflows
  • Realise and validate the outcomes from initial investigations
  • Understand the pros/cons of the infrastructure technology (Hadoop, Taverna)
  • Investigate and understand the pros/cons of the tools
    • Migration codecs, validation tools, QA tools

• Even if this image migration workflow is not used to generate preservation masters
  • Still valid as workflow for generating access copies

• TIFF to JPEG2000 migration is one possibility
  • What if there’s a problem with JP2’s or the migration software is found to have a bug
    • JP2->TIFF migration may be needed to reverse the process