Digitisation of text and images

Simon Mahony

From an original document by Claire Warwick

This document is part of a collection of presentations with a focus on electronic publishing. For full details of this and the rest of the collection see the cover sheet at: http://ucloer.eprints-hosting.org/id/eprint/34

Version 1.0

How do Resources become Digital?

- Digitisation
- Introduction to the digital image
  - History
  - Files and formats
- Scanning Techniques
- Digitising Text
- Digitisation Issues

What is digital?

•"Of or pertaining to a finger, or to the fingers or digits." OED (sv1) online

•"Of, pertaining to, or using digits; *spec.* applied to a computer which operates on data in the form of digits or similar discrete elements." OED (sv2) online.

What is digitisation?

- The process of creating a binary representation of an object that can be stored, manipulated, transmitted, and displayed, using electronic technologies.
  - 01010101010101010101000111010

- Usually used to refer to the process of sampling an object to create a digital image.

- Not a perfect copy, but a translation
  - Information can be lost and inserted

Why Digitise?

- High information content
- Significant proven public and educational benefit
  - increase resource accessibility
  - enhances ways in which contents can be studied, manipulated, or accessed
- where material is at risk
  - conservation of heavily used material?
  - existing storage medium is deteriorating?
  - possibly measure that deterioration

What to Digitise

- Printed books & journals
- Manuscripts
- Maps
- Photographs
- Transparencies
- Music manuscripts
- Woodcuts
- Line drawings
- Archaeological site plans
- Blueprints/Architectural illustrations/plans
- Medical illustrations
- Documents
- Newspapers
- Papyri and Ostraca

Resolution

- Number of horizontal & vertical pixels underlying an image
- Determined by Dots Per Inch (dpi)
- The more pixels captured, the higher the detail

Image Quality

- The higher the resolution, the higher the quality of image.
- But do you need high resolution?

What Resolution?

- 72 dpi - internet
- 96 dpi - PowerPoint/digital projection
- 150 dpi - colour lithographic printing (roughly)
- 175-225 – inkjet printing (roughly)
- 300 dpi - professional photographic print quality (roughly)
- 600 dpi - archival quality
- Best to scan at a higher resolution, then manipulate image in package
- Always save your first scan, and work from copies
- "Scan once for all purposes" – process afterwards

Formats

- Basic Data files
- BMP - Bitmap - Windows (*.bmp)
- PICT - Picture - Mac (*.pct, *.pic)
- Standard Format
- TIFF - Tagged Image (*.tif)
- standard for archival purposes
- large file sizes but no loss of data
- 600 dpi uncompressed tiff - desirable
- 300 dpi uncompressed tiff – minimum

Additional Formats

- JPEG - Joint Picture Expert Group (*.jpg or *.jpeg)
- Good for Photographs
- Loses lots of data (lossy)
- can specify quality of image
- GIF - Graphic Interchange Format (*.gif)
- Good for blocks of colour
- can specify particular colours you want to use (lossy)
- Both commonly used on the internet

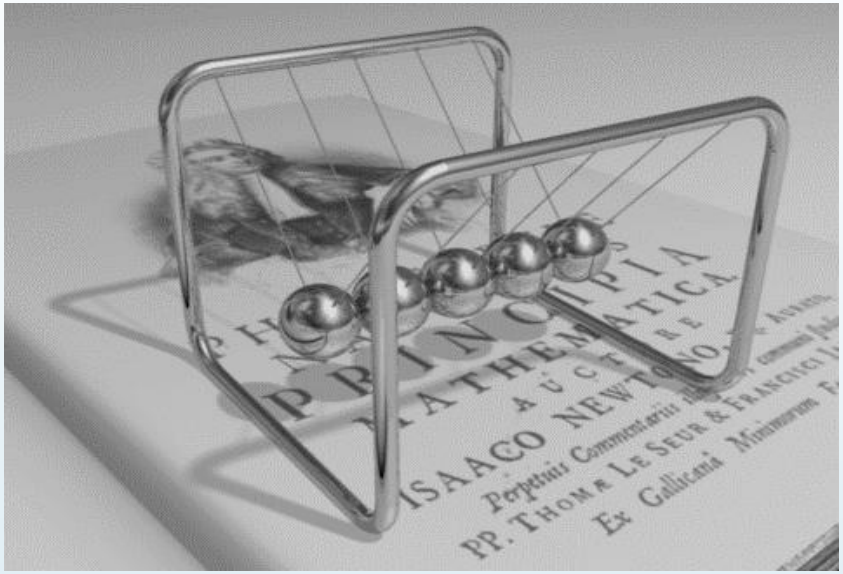
jpg / jpeg images

Images: Simon Mahony

Gif images

Simple GIF

Source: Wikimedia Commons



Animated GIF

Source: Wikimedia Commons

Scanning devices

Microfilm Scanner

Flatbed Scanner

Drum Scanner

Microfiche Scanner

Transparency Scanner

Open Book Scanner

Scanback camera

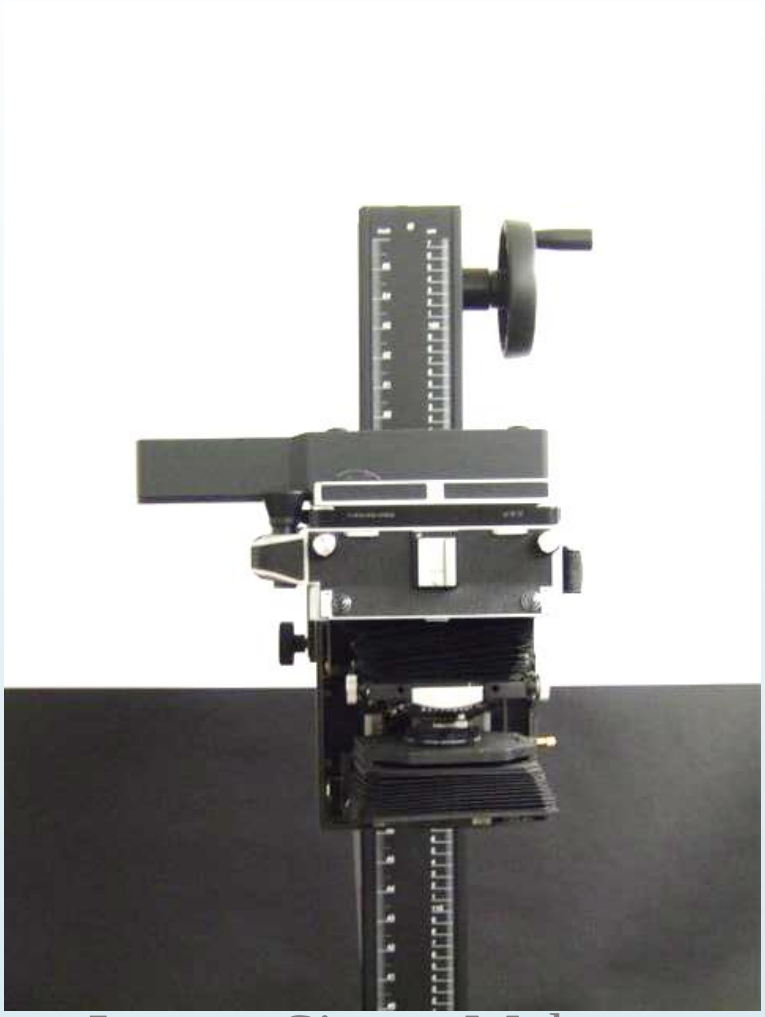


Image: Simon Mahony

Scanback camera

Image: Simon Mahony

Instant capture camera

Image: Simon Mahony

Devices- Digital Cameras

- Digital Cameras
  - Digital Cameras
  - SLR cameras with digitising backs

Post Processing

- Image optimisation vs enhancement

- Enhance the image: looks good to the eye
  - cropping / levels / colour etc

- Optimise the image: take account of the environment in which it is delivered
  - format / file size / resolution etc

Post Processing

- No general theory about image enhancement
- viewer is ultimate judge of how well a method works
- evaluation of image quality is highly subjective
- cannot control how it is viewed
- trial and error approach
- => keep a record of processes used!
- Keep a copy of your original files without processing them
- Keep a copy of metadata about original image

But...

- Time consuming and costly operation
- Does proliferation of data mean that it is harder to find information?
- Where (and who) are the users?
- Usability studies (what do users want/need)?
- Costs of maintenance?
  - Should we just re-digitise every few years as it gets cheaper? (measure deterioration of source material)

Why Digitise Text?

- Edit it
- Manipulate it
- Reproduce it
- Print it
- Search it
- Text Analysis

The Digitisation of Textual Sources

3 ways to acquire electronic text

- 1. Acquiring in electronic form (e.g. from the Internet or from an archive of electronic text)
- 2. Scanning
- 3. Keying

How Keying Works

- In-house or outsource?
  - In-house: small project; rare material that should not travel; manuscript material
- Outsourcing text
  - Can send page images rather than originals
- High accuracy levels (Up to 99.995%)
  - About one error every 20 pages
- Basic markup can be added

Advantages of Keying

- High (Up to 99.995%) rates of accuracy
  - About one error every 20 pages
- Typically, keyed by two different typists and compared by machine
- Basic textual encoding – XML or SGML – can be added, costing around 25% more
- But cost also high

When to use Keying

- If the source material is
  - Rare
  - Fragile
  - Oversized or awkward
  - Full of images, special symbols, scientific or mathematical data, or oddly formatted text
  - Handwritten or early printed book text

OCR: Optical Character Recognition

- Image of page scanned then converted into text
- Used for material that
  - Uses a clear modern typeface
  - Is clean and complete
- No smudges or tears
  - Can be fed through a sheet-feeder
  - Is formatted consistently….

OCR Limitations

- Can be a time saver, but is not perfect
- still a lot of work to convert the text to electronic form (e.g. remove page numbers, spell-check)
- Rarely more than 99.9% level of accuracy (1 error per 1,000 characters, about 10-12 lines)
- Problems with early printed books, mss, newspapers, microfilm

Text? Images? Or Both?

- Images enable use to get a sense of the original
- Often quite readable
- Often contain non linguistic information
- Appropriate for online exhibitions
- Handle special characters and illustrations

Decisions Decisions…

- Evaluate source material and format project goals
  - Who are your users? What are their needs?
- Why is the text being digitised
  - To create a copy?
  - To facilitate linguistic analysis?
- What resources are available?
  - Software, hardware, time, money
- Determine what method would be best
  - OCR or Keyboarding
- Decide how the text should be made available electronically to users
  - ASCII? HTML? PDF? SGML? XML?

Management aspects

•Assessing institutional strengths and weaknesses, timetable, and budget (Management)

•Select items from the collection to be digitised (Everything? Most Used? Cherry Picking?)

•Determining quality requirements based on document attributes (Benchmarking)

•Understanding user needs (Presentation, Delivery, Medium, Upkeep)

•Assessing long-term plans (Digital Preservation, Costs, Maintenance, Updates)

Useful links

- Library Preservation at Harvard
  - http://preserve.harvard.edu/resources/digital.html
- Cornell University Library: Moving Theory Intro Practice: Digital Imaging Tutorial
  - http://www.library.cornell.edu/preservation/tutorial/
- Technical Advisory Service for Images
  - www.tasi.ac.uk
- Now: JISC Digital Media
  - http://www.jiscdigitalmedia.ac.uk/

The National Archives

Image: Simon Mahony

The Doomsday Disc

Image: Simon Mahony

BBC Television Centre

Image: Simon Mahony

Assorted legacy videotape at the BBC Archive

Image: Simon Mahony