Material Selection for Digitisation

IMPACT Briefing Paper

IMPACT project
Neil Fitzgerald, British Library
Niall Anderson, British Library

Released under Creative Commons - Attribution-NonCommercial-ShareAlike v3 Unported (International)

Table of Contents

Briefing paper on material selection for digitsation

Executive summary

Mass digitisation involves the transformation of huge amounts of information from one form to another. The act of digitisation necessitates the creation of still more information – not only the new digital object itself, but information about what the object is, where it fits within a collection, where it is located within a digital repository, and what processes have been applied to make this digital object usable at the point of delivery. Because of the sheer volume of this information, institutions planning a mass digitisation project must establish a coherent workflow to standardise, organise, track and manage it. If the information created is to be stored long-term, then that consideration needs to be integrated into the workflow too.

A good digitisation workflow will start by establishing the nature of the content to be delivered, and the form (and functions) of its delivery. But it must also make a realistic assessment of how much of the workflow can be done within the institution itself, and how much human oversight is needed throughout the workflow in order to deliver the project on time and within budget.

Assessment and preparation of source material

Whatever the overall aim of the digitisation project, the source material will need physical assessment and sorting. It will need to be assessed to judge the status of the material, to identify any conservation issues, and to make a judgment on problems posed to digitisation by either the condition or type of material. The complexity of certain collections could mean that they are ruled out of consideration for digitisation at this stage.

Once assessed, the material will have to be sorted into size and type of item. Digitising items in batches of similar size or type cuts down on the amount of time and physical management needed at every stage of digitisation. If post-image capture processes are to be applied to the digital images (for instance, Optical Character Recognition) batches can also be usefully sorted by content.

Recommended categories for sorting by type:

- Printed text or simple line art
- Manuscripts

- Half-tone images or prints
- Continuous tone images
- Mixed (any of the above categories combined)
- Existing surrogate (such as microfilm)

Recommended categories for sorting by content:

- How much material is to be digitised in total?
- How many languages are represented in the content? Because OCR engines will only produce accurate results for languages they have been instructed to read, it is a cost and efficiency saving to process items in batches according to language
- Does the physical layout or condition of the material present problems for scanning or OCR engines?
- Do the Intellectual Property Rights of the material reside solely within the institution, or do they belong to external parties? (It may be judged too time consuming to negotiate multiple licensing contracts with external owners)
- Are the items conservation priorities? Are they intended to replace an existing physical collection as the default mode of access, or are they intended to supplement it?
- Are the items to be digitised in heavy use in their original form? If so, a strategy may have to be evolved that avoids taking them out of circulation for too long

In addition to sorting by physical type and content, an institution must also consider how, and for how long, digital access is to be provided for this material. Questions to consider are:

- What is the optimal mode of access for the material? What method of transmission will get the material to the largest number of people? If restricted access is the only present option for transmission, what technologies and methods are possible when restrictions are relaxed?
- The legal ownership of the material. If ownership does not reside entirely within the institution, it may only be able to provide limited access to the material (e.g. on institution property, or only to subscribers). If the central purpose of the individual project is to provide mass access to a large amount of information, restricted access may rule out entire collections
- What level of interaction with the content will the users be allowed to have (for example, will users be allowed to add their own comments and research as tags?)
- How does the service complement other digitised or electronic resources?

Key Terms

Continuous Tone: An image with continuous tone is one whose pigmentation – or ink – varies in concentration to the degree that the tones in the image are strong or weak, with saturated tones meaning a high concentration and light tones meaning a low concentration. The most common form of continuous-tone images are photographs.

Half Tone: A half-tone image is one where an apparently continuous image is in fact comprised of a concentration of single dots of different tones and/or colours. A common form of half-tone image is a dot matrix printout.

Image capture: In digitisation, capture is the process of creating a digital representation or image of an original through scanning or digital photography