TECHNICAL GUIDELINES FOR DIGITISATION PROJECTS

The purpose of this document is to provide technical guidelines for Wellcome Trust applicants and grantholders whose project includes the provision of funding to facilitate the digitisation of analogue material – including text, audio and video – for use over the web.

These guidelines apply to **all** Wellcome Trust grants.

215 Euston Road London NW1 2BE UK
T +44 (0)20 7611 8888 F +44 (0)20 7611 8545 www.wellcome.ac.uk

wellcome trust

TECHNICAL GUIDELINES FOR DIGITISATION PROJECTS

OVERVIEW

The Wellcome Trust best practice guidelines for digital projects falls into four areas:

1. **Creation – quality and use of standards**
2. **Preservation – sustainability and data security**
3. **Copyright – ensuring you have the right to digitise**
4. **Delivery – access and re-use**

1. CREATION OF DIGITAL CONTENT

Quality

There is no single quality measure for the creation of digital content. Different projects will require different levels of quality and different quality control measures. The key factor to consider when determining quality benchmarking is the intended purpose of the digital output. Content creators should consider both the method of capture process (the equipment to be used, consistency of lighting, position of the item, sharpness and colour management), as well as the resolution of the resulting formats, to ensure that the outputs are fit for purpose, but not over-specified.

By way of example, some projects will require very high resolution images, such as those intended to preserve highly vulnerable materials, or those with very fine detail that would be lost at a lower resolution. Other projects will not benefit from overly-high resolution images, such as those where accessing the text content is the primary use case.

Archival standards for storage

It is essential to use standards wherever possible in the creation of digital content, including metadata, to facilitate preservation and access. The recommended formats for archival storage are listed below.

Data	Recommended formats
Still images	JPEG 2000 (part 1) – lossy or lossless depending on the needs of the project TIFF (uncompressed) – where JPEG 2000 is not appropriate
Video	MPEG2 – High definition or standard definition as appropriate
Audio	WAV (uncompressed)
Catalogue metadata	ISAD(G) or EAD for archival records MARC21 for books, serials, grey literature and ephemera

Administrative metadata

All digital items (e.g. a book title, an archive folder, a manuscript, a video title) must have a unique identifier. All files associated with that item (page views, video clips, audio clips) should have unique filenames that reflect the unique identifier, plus a number sequence. File and folder names should not include spaces nor use 'special' characters such as / () & $ " etc. For example, a sequence of pages for a book with the identifier *1234* should be named *1234_0001.jp2*, *1234_0002.jp2*, *1234_0003.jp2* and so on to the final image in the sequence.

2. PRESERVATION OF DIGITAL CONTENT

Preservation refers to the ongoing management of data (files) to ensure the long-term availability of the files and their contextual information (the descriptive and administrative metadata). Long-term preservation strategies rely on the use of data standards and standardised formats to ensure that a consistent data set is created. This ensures that future management actions – including periodic validation or the migration of content held in obsolete formats – can be easily and effectively applied.

Ideally, digital content should be managed using a digital asset management system (DAM). Such systems provide a managed ecosystem supporting the ongoing management of the files and the metadata related to those files. A DAM system facilitates ongoing management actions such as file fixity checks, characterisation and format migration, as well as providing access to the content for delivery systems.

Storage associated with a DAM should be robust and secure, and either backed up or mirrored to ensure redundancy.

3. COPYRIGHT, DATA PROTECTION & LICENSING

To ensure that content can be digitised and made available it is important that grantholders pay attention to copyright and data protection issues. For general advice on these topics see the information provided by the Intellectual Property Office, the UK government and the Information Commissioner's Office.

In terms of content created under a Wellcome Trust grant we recommend that all metadata (about a digital object) is licensed under a Creative Commons Public Domain Dedication. This allows anyone to use this metadata without having to seek permission. Licence information should be included in every metadata record.

Licence information should also be provided for every image which is created. We recommend that where **no** copyright exists in the original material from where the images are sourced, the images are made available under the Creative Commons Public Domain Mark. This allows anyone to copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.

Where copyright continues to exist in the original material, grantholders should seek permission from the rights holder to make this available under the Creative Commons Attribution licence (CC-BY) or the Creative Commons Attribution, Non-commercial licence (CC-BY-NC).

Irrespective of which licence is used, **all** content digitised using Wellcome Trust funding must be made freely available on the Internet. The only permissible exception to this is when the release of such content would contravene the Data Protection Act or equivalent.

Please note that placing any Trust-funded digitised content behind a paywall/subscription barrier would be deemed to be a breach of our grant conditions.

4. DELIVERY – ACCESS AND RE-USE

Metadata

Researchers are **strongly encouraged** to make the metadata (which describes the digitised content) available in accordance with the Open Archives Initiative (OAI) protocol. This allows other organisations to harvest this content for inclusion in their systems, thus helping to ensure the maximum visibility for this content.

Content

Researchers are also **strongly encouraged** to make all digital images (which arise from our funding) available in accordance with the International Image Interoperability Framework (IIIF). This framework helps to ensure that image content is interoperable[1], which in turn will help researchers around the world find, use and re-use this content. If a researcher cannot adhere to this request, they should indicate why this is the case.

[1] A simple example demonstrates the potential of IIIF. A manuscript that has been dismembered in the past, with its leaves now scattered across various collections. If each collection which holds these leaves exposes its digitised images via the IIIF Image API, then a scholar can construct and publish a manifest that digitally recombines them to present a single coherent user experience for the manuscript in any compatible viewer.