Parliamentary Archives
Houses of Parliament
London SW1A 0PW
Telephone: (020) 7219 3074
Fax: (020) 7219 2570
E-mail: archives@parliament.uk
Web: www.parliament.uk/archives
Online catalogue: www.portcullis.parliament.uk

DIGITISATION GUIDELINES

A INTRODUCTION

Many offices and departments are considering the potential of digitisation to help them meet their business needs, whether that is providing better access to Parliamentary sources for the public, or using their own information more effectively. The Parliamentary Archives is concerned with the lifecycle of all Parliamentary records and historical information, so digitisation initiatives of any sort in Parliament need to take account of existing policies and practice in relation to both the management of current records and the digital preservation of any materials which are produced as a result.

These guidelines are for potential project managers across Parliament who are considering digitising current or historic records. Digitisation is the conversion of an analogue record (that is, a physical hard copy) to a digital copy (an electronic copy or *surrogate*) usually via scanning or another image capture process.

Creating digital surrogates of analogue records can be useful for a variety of purposes, such as:

- providing online access to information held on paper, particularly by the public
- improving searchability
- allowing previously analogue information to be accessed seamlessly with electronic records which were 'born' digital (created, saved and maintained electronically)
- reducing wear and tear on vital (i.e. critical to the House) or historical records
- reducing or - more commonly - *redistributing* the space requirement for storage of analogue records in offices
- conveying the appearance of the original artefact.

Creation of digital surrogates needs to be controlled to avoid common pitfalls. In particular, there are considerations of:

- usability
- cost
- preservability of the new digital format (and its retrieval methods)
- legal requirements, such as legal admissibility and copyright
- disposal.

These guidelines provide basic information for decision makers to help in the formulation of their plans at an early stage. Some technical information in section D provides detail on image and scanning formats. A glossary of terms is at section E.

A.1 Deciding Whether to Digitise

Valid reasons for creating digital surrogates are:

- to increase public access to useful, popular or educational material through the website or through on-demand scanning.
- to ensure that records of all the parts of the same business activity can be kept together (for example, in an environment where internal records are stored in an electronic document and records management system but external records are being received from outside Parliament in hard copy)
- to reduce wear and tear on vital or archival records
- to provide disaster contingency copies for administrative records.

Most of the other common reasons for undertaking digitisation have serious drawbacks or raise complex decisions that need to be considered before the project commences. In particular, it is worth noting that digitisation is *not* a suitable method for the long-term preservation of analogue records, and the process itself imposes its own overheads on Parliament which may outweigh the benefits. Scanning is not usually recommended for information which is to be kept for relatively short periods (for corporate retention periods for Parliamentary records see the *Authorised Record Disposal Practice* volumes).

A.2 Records Management and Digitisation

The Parliamentary Records Management Policy of April 2006 covers records *in all formats* and therefore applies to digitised versions of current and semi-current records, just as it applies to analogue copies, such as photocopies.

The functions of Parliament mean that its recordkeeping must be of the highest professional and evidential standards. It should be clear and auditable at all times:

- what exactly Parliament is relying on as a record of its business activities
- the status, primacy and relationship of each version of a record, whether electronic or analogue
- the respective degrees of reliability and authenticity of each of those versions.

The *Parliamentary Records Management Policy* and the practices agreed to implement it clarify these requirements, along with procedures of the two Houses, the standing orders, office procedures and business rules within the two administrations.

In the electronic environment, it is important to address the greater vulnerability of digital objects to malicious or unintentional alteration which may be difficult to detect. Thought also needs be given by the project manager of a digitisation project to reconciling hard and soft copy retention and following the correct procedures for deciding whether the analogue original needs to be preserved by Parliamentary Archives or whether it is required for another purpose.

A.3 Electronic Document and Records Management Systems and Digitisation

Parliament is engaged in a project examining the deployment of EDRM technologies to support the conduct of its business and maintaining its record. This is a business-led project supported by the Parliamentary Archives and PICT, aimed primarily at building the infrastructure and codifying the business rules for maintaining electronic records (for example wordprocessing, spreadsheets, and email) *throughout their life***.** It is important here to understand the difference between digitised records created from analogue sources and electronic records which have been 'born digital'.

Nevertheless, there will be areas where there are compelling reasons for digitising some analogue records to provide continuity with later electronic records, if desired, within an EDRM system. Such reasons may comprise one or a combination of the following:

- Regular, concurrent and urgent access by the House Administrations

- Business continuity.

Whether these reasons make a sufficiently strong case to offset the costs involved should be the subject of the project business case, which ought to address the alternatives and the main caveats outlined here. In particular, business cases should address the longer-term costs of securing the digitised record to ensure value for money. Expert advice on digital preservation is available from the Parliamentary Archives.

EDRM implementation might include targeted digitisation sub-projects in the future. However, it is important to bear in mind that the substantial cost of digitising and preserving both the digital surrogates *and* the original formats of these Parliamentary records cannot always be justified. The schedule for EDRM implementation can be expected to include definitive 'cut off" dates when Offices move across to keeping their records electronically, if appropriate to their business requirements.

A. 4 Copyright and Data Protection

The *Copyright, Designs and Patents Act 1988* and the *Legal Deposit Libraries Act 2003* constrain the creation and dissemination of copies of received records (ie where the copyright belongs to the party who transmitted the material to Parliament), unless they are part of the proceedings of Parliament, such as evidence accepted by a Committee. This is particularly important where it is the intention to make digitised information available over the Internet. There will also be data protection considerations where data subjects - including those whose biographical details are part of the content of such papers as well as the authors - are still alive.

The Parliamentary Archives can provide expert advice in these areas, and project managers should consult the Archives about these issues at an early stage in project planning to ensure that the work proposed does not breach this legislation.

A. 5 Costs

As well as the initial cost of the scanning itself, the digital asset produced requires an *ongoing* commitment from Parliament (particularly if digitised for public access) which may itself cost considerable amounts of money. In addition, the business case for any scanning of office files should ensure that it does not fall into the trap of justifying it on the basis of two popular myths:

1: "Scanning saves money".
In fact, the maintenance of good quality digital surrogates rarely saves money. Statistics from the United States suggests that they it can take many decades, even centuries, for the cost to dip below that of keeping the analogue version alone.

2: "Saving space".
In fact, images may take up less physical drawer space but can be huge digital files. Separate digital copies may be required for dissemination and long term archival purposes. So the costs are in reality being passed on elsewhere to the Administrations, particularly if the intention is that the images may be preserved permanently.

If after careful consideration, it is decided that digitisation is appropriate, the guidance below should be followed.

B UNDERTAKING THE WORK

The method for producing digital surrogates will vary according to why they are being created and other constraints. Considerations include:

- what are users trying to access and how?

- how many potential users are there?
- do they need to be consulted about their specific requirements?
- what is the condition of the analogue originals?
- is the scanning to be done in-house or by a contractor?
- what is the timescale for producing the digital assets?
- what resolution is required for scanning, and what impact does this have on the amount of IT storage space needed?
- does text need to be searchable?
- what is the condition and usefulness of any indices?
- what referencing system will be used for the images?
- how will that tie in with existing reference systems in place for the analogue originals?
- how long will the digital asset be required for?
- what long-term preservation requirements are there?
- how will the final product integrate with existing Parliamentary systems, particularly in the Web Centre, Libraries and Archives?
- how will it integrate with the Parliamentary network?
- taking into account all of the above, what is the cost and can it be justified?

B.1 Copies for Temporary Reference Only

Copies for short term consultation on shared drives or dissemination to workgroups by email can be created (so long as the original pages are not fragile) using desktop scanners. Provided the electronic reference versions are destroyed as soon as they have ceased to have any current value, using this procedure does not affect the management of Parliament's official record.

Attention should be paid to the settings used as these radically affect the size of the digital files created. *This is especially important if you are planning to send the scan to a number of people; it is very easy to scan to a file size of over a gigabyte and obstruct colleagues' email accounts:*

- Resolution [300*dpi* should be adequate for this purpose];
- Optical character recognition [OCR] allows the text of standard fonts to be copied and pasted;
- Colour settings: even scanning a black and white document using colour settings multiplies the size of the file generated. Colour settings may not be required even where a document has small amounts of a different colour: navy blue signatures and stamps do not normally become meaningless if rendered to black;
- File format generated: use *.pdf* or *.jpeg* format for temporary files.

Similar procedures will be promulgated in due course for early, approved EDRM adopters to convert incoming analogue material, but more robust file specifications will be prescribed.

Medium-quality images of individual documents needed for a limited period can also be taken using a digital camera. Digital cameras will produce *.jpeg* format as a default.

B.2 High Quality Surrogates

This is likely to be required for large coloured plans and pictures, using a feature-rich, open format like *.tiff* and scanning at a high resolution (600*dpi* or greater). This produces very large digital files.

Parliament will, from time to time, run large projects to digitise key content, especially for web delivery to enhance outreach. These should normally produce high-quality archival surrogates (probably in uncompressed *.tiff* format) and other formats for dissemination (see *Organising digital images* below for more on this).

B.3 Historic, Fragile or Bound Material

Fragile material must not be put through a document feeder. Bound volumes of intrinsic value in themselves (eg rare books or those with fine bindings) must not be scanned on a desktop, or industrial, scanner. An overhead digital camera and careful manual handling by trained personnel is more appropriate.

Important information may exist on the back of documents; this should be checked.

Note that the analogue originals may be destined for the Parliamentary Archives (see below).

The Parliamentary Archives can discuss the requirements and provide specialist advice on preventative and remedial conservation action.

B.4 Organising Digitised Images to Aid Retrieval and Preserve Context

Like other Parliamentary records, it is important to maintain the context of digitised images and to associate them with the functions and activities of Parliament that led to their creation or receipt. This helps with retrieval, interpretation and also assists with their disposal.

Shared drives have limited support for attaching descriptive information about the digitised object (known as metadata) and are not normally approved for the holding of Parliamentary records. They are not designed with long-term archiving in mind. They may have a limited role in holding digital surrogates of analogue material that need group access. Cross-references should be maintained to the official records either in office file lists or by adding a note in a prominent place in the drive itself.

There are a number of localised electronic document management or image management solutions on the Parliamentary estate, mostly confined to single offices. These are not approved for use with Parliamentary *records* because they do not meet several key requirements for records management. As Parliament makes further progress towards EDRM, the ability to capture metadata at source and keep born digital records digital throughout their lives will increase. Digitisation will then be an option for analogue material continuing to be received by the two administrations and in special cases.

One serious gap in many digitisation projects elsewhere is the lack of an open, exportable metadata scheme that could accompany digital files when they are exported out of the system into a future one. Major Parliamentary digitisation projects should therefore, as a matter of best practice, address in their business case how and where the digitised assets are to be managed, including the means of retrieval and creation, and metadata standards to be used. They should observe Parliamentary data standards in the interests of avoiding supplier or product 'lock-in'. These data standards are maintained by PICT as a means of ensuring best practice in systems design.

Some projects may use existing databases to perform this function, with relational links pointing to the content. Examples are PIMS in the Libraries and *Portcullis* in the Parliamentary Archives. Similar arrangements are likely for other types of planned web content, such as the Works of Art database.

C POST-DIGITISATION ISSUES

C.1 Disposal of Analogue Originals

It is the Parliamentary Archives' policy to accession archival material in the format in which it was created, wherever possible. Accordingly, offices should check the agreed disposal instruction for the type of records they intend to scan by consulting the relevant volume of the *Authorised Record Disposal Practice*.

Any analogue record with the following disposal instructions:

"*Transfer to the Parliamentary Archives*"

"*Contact the Parliamentary Archives for appraisal decision*"

MUST be sent to the Parliamentary Archives and *NOT* destroyed when digital surrogates are created.

C.2 Disposal of Digital Surrogates

As a general rule, a digital surrogate will only be used to increase the usability, accessibility and longevity of an analogue original. If a digital surrogate is made of a record that will be destroyed in the fullness of time according to the Disposal Practice, the surrogate must be destroyed in accordance with that instruction.

If the correct procedures suggested above for managing analogue records and their digital surrogates have been followed, it will not present any significant problem to ensure that the correct retention period has been applied when final disposal falls due.

In exceptional circumstances, it may be agreed that a surrogate may outlive the analogue original. This is counter to normal Parliamentary practice. It is strongly recommended that this scenario and alternatives are discussed *in advance* with Parliamentary Archives: it may be that physical conservation of the analogue version, with the creation of a temporary digital surrogate are a more appropriate strategy.

C.3 Preserving Digital Surrogates

Long-term preservation of digital images for educational or research use requires consideration within the broader digital preservation strategy. As with 'born digital' records, the preservation of scanned images involves addressing twin challenges from creation:

- Secure, recoverable storage
- Countering the threat of technical obsolescence.

A digital preservation strategy for Parliament is being developed by the Parliamentary Archives with key stakeholders in the Libraries and PICT at the time of writing. This will outline the principles to be followed to maintain access to digital archival material over time, even permanently.

As this guidance reveals, there are a series of linked issues that need to be addressed if the benefits of targeted digitisation are to be realised without the pitfalls. Digitisation alone is not a sustainable records preservation strategy. The digital surrogates themselves will require preservation and it may not be appropriate to destroy the analogue originals, either for business or legal reasons. In that scenario, controlled copying through digitisation may be appropriate, though steps must be taken to clarify on how (i.e. in which manifestation, analogue or digital) Parliament is maintaining its record of the original business activity. In addition uncontrolled copies could cause problems with disposal management and with the answering of Freedom of Infomation, Data Protection and Environmental Infomation Regulations requests.

D SOME TECHNICAL INFORMATION

D.1 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); and
- *raster* images 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

D.2 Optical character recognition (OCR)

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!

E GLOSSARY OF TERMS

Analogue A physical object (eg book, document, photograph), rather than its electronic manifestation.

EDRM **Electronic Document and Records Management**. A networked system which acts as an electronic filing cabinet capable of keeping corporate documents and records securely, and disposing of the latter automatically at the appropriate time.

dpi	**Dots per inch**. A measure of the **resolution** of the image.
Metadata	'Data about data', that is, a description of a digital asset to aid searching, retrieval and preservation. There are various international metadata standards required to describe a digital asset comprehensively.
OCR	**Optical Character Recognition.** A method of electronically analysing **analogue** text (eg from a printed or manuscript document) and converting it into electronic text for searching and manipulation. See also section D.
JPG	An image format widely used for images in web pages, and for digital photographs. See also section D.
PDF	**Portable Document Format.** A software format created by the software company Adobe, into which many other proprietary 2D and 3D document formats (eg Microsoft .doc or .xls) can be converted for reading via a single viewer. Conversion to PDF enables users to view documents originating in a software they do not use, so long as they have Acrobat Reader (the viewer) installed on their computer. See also section D.
Resolution	The complexity of an electronic image, dependent on the number of pixels (individual colour units) created during the scanning process.
Scanning	The process of capturing an electronic image from an **analogue** original. Scanning can be combined with **OCR** to create an image of a document as well as a copy of the text of that document.
Surrogate	The term used by archives and libraries for a copy of a record or book, particularly one created to prevent wear and tear on the original. A surrogate can be **analogue** (eg photocopy or microfilm) or electronic (eg scanned image).
TIFF	The standard image format for scanned images. See also Section D.

MARCH 2008