A Code of Practice For Legal Admissibility of Information Stored on Electronic Document Management Systems

Table of Contents

Background

This publication has been developed with the assistance of a group of leading UK companies, consultants and associations, as a recommended Code of Practice for legal admissibility of information stored on electronic document management systems.

This Code of Practice has been prepared by the British Standards Institution technical committee BSFD/14 and is the result of amalgamation of work from the following bodies, represented on BSFD/14:

Document Management Forum (DMF)
Image and Document Management Association (IDMA)
Legal Images Initiative consortium (LII)

United Kingdom Association for Information and Image Management (UKAIM)

BSFD/1 4 would like to thank representatives of the following authoritative bodies for their help and comment on the final draft:

Centre for Commercial Law Studies
Law Society of Scotland
Lord Chancellor's Department
The Data Protection Registrar
The Department of Trade and Industry (DTI)
The Government Centre for Information Technology (CCTA)

Acknowledgements

The Editor would like to thank representatives of the following companies, consultancies and associations for their support and assistance in the development of this Code of Practice.

3M (UK) PLC
Association of Computer Telephone Integration Users and Suppliers (ACTIUS)
Advent Imaging Limited
Ashford Borough Council
Auto-trol Technology Limited
Bell and Howell
Bird and Bird
Blueprint
Brighton University
CCTA
Centre for Commercial Law Studies (University of London)
Cimtech Limited
Document Image Technology
Document Managers Forum (Computing Suppliers Federation) (DMF/CSF)
European Electronic Messaging Association (EEMA)
Fujitsu Europe Limited
Headway Computer Products

Headway Technology Group
Hewlett-Packard Limited
Image and Document Management Association (IDMA)
Imtec Group
Intergraph (UK; Limited
JTS Systems
Legal Images Initiative consortium (LII)
Lloyds Bank PLC
Lombard Document Systems Limited
London Borough of Enfield
London Transport
Marc Fresko
Maxoptix Europe Limited
Microgen (UK) Limited
MR Data Management Group PLC
National Westminster Bank
North Hampshire Hospital
Oki Systems (UK) Limited
Q Star Limited
Records Management Society of Great Britain (RMS)
Scottish Nuclear Limited
Sony (UK) Limited
SSI Microcad
Tekdata Limited -
Trimco Enterprises Limited
UK Banks Credit Card Committee
United Kingdom Association of Information and Image Management (UKAIM) Xerox Imaging Systems

Detailed Document Breakdown

Introduction

The production and storage of documents on computer systems has become common practice. It is therefore inevitable that these stored documents will increasingly be used in their electronic form as a basis for business transactions, and will be produced, transmitted, and stored in significant numbers.

There has been much discussion about the value of documents stored on document management systems when documents are required to be kept as evidence for a considerable time. It is crucial that a discipline is commonly agreed so that the value of these documents as evidence can be maximised. It has not been possible at this time to produce a set of requirements. Instead, a Code of Practice has been developed which will evolve as the technology and electronic commercial practices mature.

This Code of Practice is the result of the merging of work from two organisations, namely the Legal Images Initiative (formed by the Image and Document Management Association) and the Document Management Forum (a group of the Computing Suppliers Federation).

A document entitled 'Principles of Good Practice for Information Management' [1] written by two of the authors of this document, contains a detailed explanation of the background to each of the sections in this Code.

This Code of Practice is to be used as a basic reference document. The Code covers data files stored on Write-Once-Read-Many (WORM) optical storage systems. As such it covers WORM, multi-function media Systems used in a write once mode, and compact disk recordable (CD-R) Systems.

This Code of Practice does not guarantee legal admissibility. It seeks to define the current interpretation of best practice.

This Code of Practice covers issues such as system planning, implementation, initial loading, and the procedures for the use of the system. It pays particular attention to setting up authorised procedures and subsequently the ability to demonstrate, in a Court of Law, that these procedures have been followed.

This Code of Practice defines essential procedures to be implemented in order to conform to the Code. It does not follow that documents held on systems that do not conform to all the essential processes and procedures in this Code are not legally acceptable. However, it is likely that it will be more difficult to prove their integrity in a court of law.

A number of terms used in this document are defined in **1.3.**

Weight of Evidence and Document Destruction

It is important to determine, in advance, how a document would be presented to a court of law, and if weight of evidence or courtroom tactics could be unduly influenced by the destruction of the original document, by the document storage system or by the access control systems.

It may not always be possible to give a definitive recommendation regarding the destruction of original documents. Until there is a request to produce a document, the reason for the request may not be known, it is the reason for the request that will indicate whether, if possible, the original document should be produced. The Company Solicitor will be able to provide a view as to which types of documents are most likely to be disputed regarding their authenticity rather than their content, There may be different considerations for civil and criminal law. In a criminal case, the prosecution faces a much higher burden of proof 'beyond reasonable doubt' than in civil proceedings 'on the balance of probabilities.

Authenticity

It is important to be able to demonstrate that the computer system has been functioning properly (i-e. according to agreed procedures) in order to authenticate documents stored on the system. Documents may be rejected if this cannot be shown.

In most cases, arguments are over what a document says rather than the authenticity of the document. However, the adversarial legal process means that the other party may try to discredit evidence.

Arguments over admissibility of evidence can lead to an investigation into the system that produced the paper, and the method of storage, operation and access control, and even to the computer programs and source code. It may be necessary to satisfy the court that the information is stored in a proper manner. This could be a tactic used to try to discredit the evidence and to make inadmissible that and any similarly stored documents that are produced. Hardware reliability, for example, could be used to discredit the document storage system. It could enable the whole system to be questioned and documents stored within it ruled inadmissible.

Documented procedures for storage, maintenance and auditing access to the documents will minimise this risk.

Photocopies, microfilm and image processing

Image processed documents will be treated as secondary evidence in the same manner as a photocopy or a microfilm image. Photocopies and microfilm images are admissible as evidence. Some photocopies use a raster scan copying mechanism, essentially the same as an image processing scanner It follows that document image processed documents are likely to be admissible, with the same weight of evidence as photocopies and microfilm images, although this has yet to be tested in courts.

Document storage

No matter how an organisation stores business documents, it is the responsibility of the executives of the organisation to be able to produce the document (or a copy) when required. The company secretary and the manager of the document storage systems are responsible for this document retrieval process, and not the vendor of the storage system. Thus, the advice of the company secretary should always be sought before implementing any document storage system, particularly when the original documents are subsequently destroyed.

The procedures by which documents are stored and accessed are vital in satisfying a court of law about the authenticity of a 'copy' of a document, and the inability to tamper with it. All copies of documents (photocopy, microfilm or image processing) will be treated by a court of law as secondary evidence, with a subsequent reduction of weight of evidence, if the authenticity of the copy can be questioned. For example where the content of a document is under question, the original or a copy should be treated with equal weight, but if a signature is being disputed, then the original document is likely to carry more weight than the copy.

There may be some confusion about 'originals' and 'copies'. Many items to be scanned are actually themselves photocopies. The original document may reside in a file elsewhere. It may be necessary for the image processing system to indicate whether an image was from the original or from a copy of it.

Storage and access procedures

Because of the duration of storage of many documents, the person who 'certified' a system, or a document stored on it, may not be able to give evidence in person. It is essential that a proper system for auditing and certifying is implemented to demonstrate that the integrity of the system has been maintained from the time the document was stored.

Regular audits of the system should be performed, and certificates obtained from the company auditors. This is in line with current procedures for microfilmed documents. Although formal affidavits will not usually be necessary, advice should be sought from the company solicitor, particularly if the original documents are to be destroyed.

It may help demonstrate the proper functioning of a system if a copy of the audit record is stored in the image system at the time of audit.

As well as the specific details included in this Code, users should also comply with the relevant sections of the following British Standards:

BS 7768: 1994 - Management of optical disk (WORM) systems for the recording of documents that may be required as evidence.

BS 7799: 1995 - Code of Practice for Information Security Management.

Of major importance to this Code is the Civil Evidence Act 1995 [2] The Act introduces a flexible system whereby all documents and copy documents, including computer records, can be admitted as evidence in civil proceedings. The court judge will still have to be persuaded to treat that evidence as reliable, and so organisations will have to put in place procedures to prove the authenticity and reliability of the record.

Sections 8 and 9 of this Act address the hub of the issue:

8 (1) Where a statement contained in a document is admissible as evidence in civil proceedings; it may be proved:-

(a) By the production of that document, or

(b) Whether or not that document is still in existence, by the production of a copy of that document or of the material part of it,

authenticated in such a manner as the court may approve.

(2) It is immaterial for this purpose how many removes there are between a copy and the original.

9 (1) A document which is shown to form part of the records of a business or public authority may be received in evidence in civil proceedings without further proof.

(2) A document shall be taken to form part of the records of a business or public authority if there is produced to a court a certificate to that effect signed by an officer of the business or authority to which the records belong.."

Similar work is being progressed by the Home Office on a Police and Criminal Law amendment.

Format of code

The Code is divided into five parts (see **2** to **6**), each of which contains details of processes and procedures that need to be put into place to ensure conformity with this Code.

A detailed explanation of the background to each of these parts, is given in the 'Principles of Good Practice for Information Management' [I]

Section 1 General

1.1 Scope

This Code of Practice describes the use of electronic document management systems to store documents, where the issues of legal admissibility, authenticity and evidential weight of information contained in these stored documents is important.

The Code is for use with document management systems incorporating write once optical media as the storage device. As such it covers WORM, multi-function media systems used in a write once mode, and compact disk recordable (CD-R) systems.

The Code does not cover document management Systems that incorporate re-writable media (for example magnetic storage), as the controls necessary are outside the scope of this document. This does not mean that such systems cannot be used for the storage of documents that may be required as evidence. However, such systems require more stringent controls than those described in this document.

The Code covers any type of data file controlled by the document management system. Data files may potentially contain text, image, CAD data, moving and still video images, and audio, or any combination of these or similar data types.

Data files may be created by the document management system, or may be imported into it. The Code covers all such data files, either created and/or imported directly or through a network system, from the time at which the system assumes complete control of the data file. Such networks may be local or wide area.

While the Code covers aspects of document management that impinge upon the issue of legal admissibility of digitised images, it also covers aspects that may affect the use of the images in a legal context, even where admissibility per se is not at issue. Such aspects include the legibility and completeness of the document images, and the transfer of the images into other systems.

The code covers the capture of digitised images both from original documents and from microform versions of the original documents. In the latter case, users should be aware of the implications of the processes used in the microfilming of the original documents.

The Code is intended for:

Systems Integrators whose equipment provides facilities to meet the requirements of end users.

End users who wish to ensure that the digitised images of documents captured by and stored on such systems may be used with confidence as evidence in a Court of Law.

Where users wish to claim adherence to this Code, paragraphs identified by text in bold type are considered essential in so far as they apply to the specific application concerned. Other paragraphs contain recommendations that should be followed where practical.

1.2 References

1.2.1 Normative references

This Code of Practice incorporates, by dated reference, provisions from another publication. This reference is made at the appropriate place in the text and the cited publication listed in Annex D. For this dated reference, only the edition cited applies; any subsequent amendments to or revisions of the cited publication apply to this Code of Practice only when incorporated in the reference by amendment or revision.

1.2.2 Informative references

This Code of Practice refers to other publications that provide information or guidance. Editions of these publications current at the time of issue of this Code of Practice are listed in Annex D. but reference should be made to the latest editions.

1.3 Definitions

For the purposes of this document the following definitions apply.

1.3.1 Compact Disk Recordable (CD-R)

Optical disk conforming to the specification of a compact disk (CD), which can be written to only once (or in multi-session disks only once per session) by a user write system.

1.3.2 Compression ratio

Ratio between the number of bits in a digitised image before compression and that after compression.

1.3.3 Computer Output to Laser Disk (COLD)

A computer system which includes the storage and retrieval on optical disk of text-based alphanumeric computer output.

1.3.4 Critical records

Records which are fundamental to the functioning of an organisation.

1.15 Data file

A single record or collection of data records that are logically related to each other, and are handled as a unit.

1.3.6 Deletion

The process of logically removing a document from a system, often by deleting an index reference. In this case, it is (technically) possible to undelete the document (see also **1.3.10** expungement.)

1.3.7 Digital, digitised Image

Image composed of discrete pixels of digitally quantized brightness.

1.3.8 Digital signature

A data block appended to a data file such that the recipient of the data file can authenticate the data file, and/or can prove that it could only have originated from the purported sender.

1.3.9 Edge enhancement

Technique for sharpening the appearance of the boundary between black and white on a digitised image.

1.3.10 Expungement

The process of removing a document from a system, and leaving no evidence of the document ever having appeared on the system.

1.1.11 Forms removal

System (usually software) which removes a 'fixed' overlay from a digitised image, leaving the variable data only.

1.3.12 Grey scale image

Image formed of pixels containing grey scale information.

1.3.13 Group (x) compression

A file compression system for digital data (see references). Where (x) is either 3 or 4.

1.3.14 Hierarchical storage system

A data file storage system using a variety of storage devices, from high cost, fast access devices to low cost, slow access devices, and that data files can be passed under system control from one device to another

1.3. 15 Intelligent Character Recognition (ICR)

Similar to OCR, but using sophisticated software techniques to recognise characters from the way they are formed (as opposed to comparing 'unknown' characters with standard pre-stored patterns) and from their context.

1.3.16 Integrated Services Digital Network (ISDN)

Switched digital transmission service for voice and/or data.

1.1.17 JBIG

An ISO Standard for lossless compression techniques for bi-level images. This system offers improved compression over Group 4 compression (typically 40% to 80%). For half tone material (e.g.. newspapers and magazines) compression ratios are superior to those obtained by Group 4 methods.

1.1.18 JPEG

An ISO standard for compression of grey scale and full colour images, developed by ISO and CCITT's Joint Photographic Expert Group. JPEG compression is usually used in 'lossy' compression mode, but a lossless mode is also available.

1.3.19 Lossless compression

Compression technique where the decompressed image is identical to the original uncompressed image (i.e. no information is removed during the compression and decompression processes.)

1.3.20 Lossy compression

Compression technique in which information in an image to which the eye is relatively insensitive is removed. High compression ratios (e.g. around 50:1) can be obtained with little observable image degradation. Lossy compression necessarily means that the decompressed image may not contain all of the information contained in the original image.

1.3.21 Multi-function optical system

Optical disk system which can use both WORM and erasable/rewritable optical media.

1.3.22 Optical Character Recognition (OCR)

Technique for the recognition of characters from a digital image.

1.3.23 Pixel

Smallest element of a digital image (from 'picture element').

1.3.24 Public Switched Telephone Network (PSTN)

A telephone switching system used by the public for voice and data transmission processes.

1.3.25 Scanning

Operation which converts the image of a document into a digital form, by detecting the amount of light reflected from elements of a document - one line at a time.

1.3.26 System files

Computer readable files held on a computer for use in the control and operation of the system.

1.3.27 Write-Once-Read-Many (WORM)

Type of Optical Disk where each logical sector may be recorded only once, but which can be read may times.

Section 2. Representation of information

2.1 General

Information is one of the most important assets that any organisation has at its disposal. Everything an organisation does involves using information in some way. The quantity of information can be vast, and there are many different forms of representing and storing it. The value of information used and the manner in which it is applied and moved within and between organisations may determine the success or failure of that organisation.

Information, like any other resource, needs to be classified, structured, validated, valued, secured, monitored, measured and managed efficiently and effectively.

Information that is to be stored on a document management system needs to be classified according to its life cycle. This classification should then be used to determine the storage procedures necessary.

Document life cycles include:

- Creation
- Retention period

- Access
- Revisions
- Destruction

Once classified, the access and retrieval procedures should be defined.

The medium on which, or in which, the document is stored needs to be specified. Each of these has different longterm storage characteristics. The three forms which most organisations need to recognise and address are:

- ½ Paper
- Microform
- Electronic

2.1 Policy document

A policy document should be produced, dealing with policy on:

- Form in which documents art held (see 2.1)
- Document life cycles (see 2.1)
- Legal advice sought and acted upon

2.2 Policy document

A policy document should be produced, dealing with policy on:

- **Form in which documents are held (see 2.1)**
- **Document life cycles (see 2.1)**
- **Legal advice sought and acted upon**

Section 3. Duty of care

3.1 General

It is essential that an organisation is aware of the value of information that it stores, and execute its responsibility to that information under the 'Duty of care principle.

To fulfill this objective, tile organisation must:

- Be aware of legislation and regulatory bodies pertinent to its industry.

- Establish a certificate of accountability and assign responsibility for activities involving electronic document management at all levels.

- Keep abreast of developments by keeping in contact with the appropriate bodies and organisations.

An organisation should have appropriate levels of security for managing its information agreed and documented.

BS 7799: 1995, *A Code Of Practice for Information Security Management,* should be used as a basis for the implementation of these security procedures.

It is essential that systems are adequately managed. To comply with this requirement it is essential that the relevant sections of the following British Standard are implemented:

BS 7768: 1994, *Recommendations for management of optical disk (WORM) systems for the recording of documents that may be required as evidence.*

The objectives of this Code are to provide:

- A common basis for companies to develop, implement and control effective security management practices;

- Confidence in inter-company trading.

The Code can be used as a common reference standard for inter-company trading and for sub-contracting or procurement of information technology services of products.

3.2 Consultations

The implications of installing an electronic document management system to store documents that have legal significance are far reaching. Such systems are becoming common, with various codes established by organisations involved with these systems.

It is essential to consult with interested third parties at the planning stage and before the system is installed.

Consultations should take place under the following topics:

- Legal issues (civil law and/or company law - contracts and disputes.)
- Government bodies.
- Special regulations.

Each organisation should determine the levels to which these consultations will be made:

- International law
- National law
- Community
- Industry sector
- Organisation
- Department
- Individual

Consultation with the following organisations should be considered:

- Bank of England
- Civil Aviation Authority
- Food and Drug Administration
- ElM Customs and Excise
- Inland Revenue

The following organisations or individuals should be consulted:

- Company secretary
- Company legal department
- External legal advisors
- Internal audit group
- Company auditors
- Government bodies

For example: Ministry of Agriculture for documents relating to the production of food Public Records Office Ileab and Safety Executive

- CCTA
- Industry regulators

Section 4. Business procedures and processes

4.1 General

This section deals with the operating procedures a user should implement. It should be possible to demonstrate to external parties such as auditors or lawyers that the system conforms to the Code at the appropriate times. These procedures should be documented in the user manual.

The procedures that should be included within this user manual are described in **4.3** to **4.19.**

4.2 User manual

The user organisation should develop its own user manual for the document management system. Where the user organisation operates a quality system (such as BS EN ISO 9000), then the user manual should be included in the quality system. Suitable training should be given to all staff who operate the system, to ensure that the procedures detailed in the user manual are adhered to. This user manual, in addition to any vendor supplied manuals for the system, should include the following topics:

- Document types
- Preparation of documents prior to scanning
- Photocopies
- Batch control
- Scanning processes
- Scanning specific documents
- Image processing
- Compression Techniques
- Indexing
- Quality control
- Output processes
- Document retention
- System maintenance
- Security and Protection
- Backup and data recovery
- Use of bureau services
- Remote transmission of data files
- Voice files
- Document status change

An annual (or more frequent) system audit should be carried out to ensure that any changes to the documented procedures still meet the operational requirements, and the requirements of this Code. This review should be certified by the person responsible for the operation of the system.

Procedures should be included in the user manual to confirm that the procedures in the manual are being adhered to.

4.3 Document types

Data files of three types are envisaged, each of which may require different procedures:

- Generated by a computer system - also known as encoded data files
- Scanned images / digitised voice and /or video

- Generated at a remote user or third party site, in either of the above two types

For each type, the validity of the source data file and its committal to optical media needs to be controlled, by checking the origin of the data *files,* and conforming to the procedures documented in the user manual.

Where the data file is of a compound nature (e.g. forms and data or image with voice annotation), the linkages and relationships between them should be treated in the same way as the data files themselves.

4.4 Preparation of documents prior to scanning

Documents should be examined prior to the scanning process, to ensure their suitability. Such factors as their physical state (thin paper, creased, stapled, etc.), and the attributes of the information (black and white, colour, tonal range, etc.) should be noted. Procedures for this examination process should be documented in the user manual.

Documents that may be adversely affected by the scanning process (e.g. damaged or delicate documents) may be photocopied before scanning.

Documents containing paper or ink colours that do not produce legible scanned images may be photocopied before scanning.

Photocopiers may respond to different colours than scanners. Photocopies should be examined to ensure that information is not lost during this process. It is only in very exceptional cases that this technique does not produce satisfactory results.

Where there are substantial contrast or density variations over the area of the original and the results of scanning from the original are unsatisfactory, photocopies may be made if this demonstrably improves the image quality.

Care should be taken when multi-page documents use staples or clips to join a number of pages together. These should be removed in such a way as to ensure that no damage is caused to the original that may affect the capture of the information from the document. All pages should be kept together and in the correct order before, during and after scanning.

With folded documents that are too large to be scanned as a single full-sized image, photo-reductions may be made which are then scanned, and/or multiple scanned images may be captured from the original or from photocopies thereof.

If photo-reductions are made checks should be made to ensure that there is no loss of detail in the scanned images compared to the original caused by the effective resolution of the image (compared to the original) being reduced.

If multiple images are captured, these should be overlapped to ensure no loss of information at the edges between adjoining images.

Documents should be grouped into 'batches'. The definition of a 'batch' will be application dependent. For example, if the documents are in file covers, and the average number of documents per file cover is relatively large (e.g. about 100 pages), then the file cover itself

may constitute the batch. If the file covers contain relatively few documents (e.g. less than 10 pages, average), then a batch could consist of more than one file cover.
If the documents are on roll microfilm, the film roll may be the batch. The definition of a batch is decided on the basis of convenience. The average number of documents per batch should not be so big as to be difficult to manage (e.g. thousands of documents per batch), not so small that sampling on a batch basis would result in a zero sample for a batch.

For some applications, a 'batch' may not be easily defined. In these cases, a batch may be defined as 'those documents input during a specified time period'. Thus, for example, a batch could be all documents input during a day or an hour

4.5 Photocopies

Some documents may need to be photocopied prior to the scanning process. The procedures used for this photocopying should be documented in the user manual.

Where a photocopy is made of an original document as part of the preparation procedures, such action should be recorded and a record that the image has been captured from a photocopy of the original should also be stored with the document image (i.e. as an associated data file).

Where the document to be converted is itself a photocopy, this should be recorded, whether or not the document is photocopied again as part of the preparation procedures. This is to ensure that an image may be correctly identified as a true facsimile of an original document, even if an intermediate photocopy has been taken as part of the preparation procedures, and to distinguish such images from images of photocopies made under unknown conditions.

Where the source document for the conversion is known to be a photocopy, this may be recorded by stamping the document with the word 'photocopy'.

Where it is not known whether a document is an original or a photocopy, it should be treated as the original document.

4.6 Scanning processes

The user manual should include details of the operational procedures used in the scanning processes, taking into consideration application and imaging system characteristics.

Records should be kept in the system audit trail of key information concerning documents that have been imported into the system. These records may be kept manually or automatically or a mixture of both.

Records should be subject to 'at least as good as' the normal internal records management procedures that are used for other 'critical records' in the organisation.

The system should give each document a unique identity that cannot be changed

or removed, except as described in clause 5.13 on expungement. This unique identity could be a system generated sequence number, which could be used for internal control purposes only.

Information held in the records should include as a minimum the following:

- Unique identifier for each batch of documents
- Date and time of scanning
- Person who performed the scanning
- Type of material scanned, e.g. paper documents, roll microfilm, aperture cards
- Number of documents scanned, e.g. number of documents, number of pages, number of microfilm frames
- Details of 'post-scanning processing' performed

Records should be kept on a batch basis, so that it is easy to check:

- That all required activity has been performed for that batch;
- □ That a note of any anomalies or discrepancies has been made (e.g; number of pages written to optical disk not agreeing with number of pages scanned);
- That appropriate quality control procedures have been completed successfully;
- That records of appropriate exception processing have been made;

The information held in these records should be written to the audit trail.

If the paper document scanning is to be done using an automatic document feeder (ADF) attention should be paid to the risk of multiple documents being fed inadvertently. To detect this it is advisable to pre-index the documents in order to generate a sheet count, or to use a batch header procedure.

This sheet count can subsequently be compared with the scanned page count; any shortfall will indicate either that more than one page has been fed at once or that a page has been misplaced between pre-indexing and scanning.

If a simplex scanner (i.e. one that scans only one side of a document at a time) is used to scan double-sided documents, care should be taken to ensure that every double-sided document is reversed and the other side scanned.

4.7 Scanning specific documents

Different types of documents require different scanning characteristics.

Any variations in the operation of the document management system due to the type of document being scanned should be detailed in the user manual.

4.8 Image processing

Image processing techniques can be used to improve the quality of an image. If these procedures are optional, then their use should be described in the user manual.

4.9 Indexing

Indexing is a vital part of the process of storing documents on optical media. Should indexing information be lost, then the related documents may also be lost.

Indexing can be either automatic (i.e. performed by the system without operator intervention), or manual. If manual indexing is performed, care should be taken to follow the procedures documented in the user manual.

Procedures for indexing documents should be described in the user manual. These procedures should include methods for checking the accuracy of the index records created.

Automatic indexing rules and processes should be described in the user manual.

A copy of the index file should be stored on the same optical disk as the documents to which it refers. Sufficient space should be left on the optical disk for updated index files that may be required as part of a document deletion process. If insufficient space is available for a new index file, then a procedure for either the use of additional optical disks for the new index files, or the copying of all of the data files onto new optical disks together with the new index files, should be documented,

Some systems allow some index information to be stored when the document is captured. This may then be combined with additional manual index entries at a later time. This is acceptable provided the procedures and processes of this Code are observed.

The system should record in the audit trail the creation of and amendment to all indexes. Information about the date and time of each creation or amendment should be recorded in the audit trail. If an index record is being amended, details of the change should be recorded in the audit trail.

Operators should be trained how to ensure, on a regular basis, that the processes are running correctly, and how to handle documents that have not been indexed because of errors. Procedures should be in place for the amendment and correction of these indexing errors.

Indexing process may include procedures for the detection of missing images. Indexing from displayed images will not detect missing images unless the displayed images are checked against the originals, or there is a defined sequence of documents (for example by sequential document numbering).

Quality control criteria for index data accuracy levels should be realistic given the method used for index data capture, the typical random error criteria achieved by data entry personnel, and the legibility of the source material.

Index data accuracy criteria may vary depending upon the application. In some cases the accuracy may be defined as the maximum acceptable number of characters in error per 1000 characters captured (or percentage equivalent). In other cases the accuracy may be defined as the maximum acceptable number of words (or similar cluster of characters, for a customer or part number) containing any error (whether of one or more characters).

4.10 Quality control

The description below summarises some of the main aspects to quality control.

A sample set of original documents, or of documents equivalent in characteristics to the original documents, should be assembled for the purposes of benchmarking scanning system performance against the quality control criteria.

The documents in the sample set should be representative of the complete set of documents that are to be processed. They should include examples of source documents whose quality is poor relative to those of the majority of the documents.

Care should be taken when evaluating the results of a quality control procedure. Results obtained may depend upon the specific output device (e.g. visual display unit (VDU) or printer). The procedures documented in the User manual should specify the evaluating process, bearing in mind the output device used.

If a printer is to be used for quality control procedures, if possible the printer resolution should be equal to or greater than the resolution of the scanned images and should be capable of accurate reproduction of grey scale or colour in applications where this is relevant. This is to ensure that scanner operating staff, quality control staff and end users are aware of what is practically possible as regards scanned image quality.

Quality control criteria should be set for scanned image quality. The criteria used should be agreed by all parties likely to be affected by image quality (e.g. in-house or out-of-house users).

The criteria may cover overall legibility, smallest detail legibly captured (e.g. smallest type size for text), completeness of detail (e.g. acceptability of broken characters, missing segments of lines), dimensional accuracy compared to original, scanner-generated speckle (i.e. speckle not present on the original), completeness of overall image area (i.e. missing information at the edges of the image area), density of solid 'black' areas, and colour fidelity.

Quality control procedures should be undertaken by personnel other than those responsible for the scanning of the material being examined.

Quality control procedures should be related to the batch process as defined earlier, enabling acceptance or rejection of such a batch independently of any other batch.

Where the quality control procedures involve sampling of the scanned images and related text data, the proportion sampled need not be fixed but may vary from time to time depending on the frequency of problems encountered or the nature of the source material, will not normally be practicable to check all processed material and generally only a proportion of the material processed will be checked.

For example, when starting scanning initially a relatively large sample may be selected (e.g. 20%), which may be reduced (e.g. to 10% or even 5%) as the users become confident that the required quality standards are being met consistently.

Sample sizes should be determined in accordance with BS 6001.

Quality control criteria for image quality should be realistic given the nature of the source material and the characteristics of the scanning equipment. The criteria should be based upon the sample set described at the beginning of this clause.

The quality control procedures should be used periodically to check that the system performance remains stable. Hard copy prints should be made of the scanned images of the test targets and compared to the test targets themselves to determine whether the quality criteria are met, as described in the procedures.

Quality control check frequency may vary depending on the system usage. The frequency should be related to the expectation of deterioration in system performance, which may require recommendations from the system supplier and also experience in the use of the system. Initially, it may be appropriate to scan a test target every few thousand pages scanned.

If double-sided (duplex scanners are used, double-sided test targets should preferably be used. Single-sided test targets should only be used with duplex scanners if double-sided test targets cannot be obtained.

Test targets are not representative of the documents actually being scanned and should not he regarded as a substitute for the sample set of documents.

The results of all quality control checks, including Test Target scans, should be recorded in the quality control log.

4.11 Output procedures

A description of the procedures for the creation of copies of documents held on a document management system should be included in the user manual. This creation may be in the form of a physical document, or may be a screen display.

Where a document is produced, these procedures should include the use of an authorised signature to authenticate the output from the system where appropriate.

Where a display of a document on a screen is produced, evidence of the authentication of the display should be provided. The procedures for this authentication may vary depending upon the equipment used to produce the display.

4.12 Document retention

Procedures for the retention or destruction of documents should be described in the user manual.

Even where it is intended that all original documents are not retained after capture of digitised images, there are some instances where documents should in any case be retained. These include the following situations:

- **Where a photocopy or photocopies of the original have been made to aid the scanning processes, and the digitized image captured from the photocopy(ies), both the original and the photocopy(ies) should be retained. This avoids any risk of rejection of an image on the grounds that it is not a facsimile of the original document.**

- **When the original is of poor quality and recorded as such in the index data to the digitised image. This deals with the possibility of it being suggested that an image was deliberately made illegible.**

- **Where an original contains physical amendments that cannot be identified as such on a scanned image.**

No original source documents should be destroyed until the write processes have been verified and appropriate backup procedures completed.

4.13 System maintenance

The document management system should be maintained by qualified personnel to ensure that its performance, specifically as regards quality, does not deteriorate. A maintenance log should be kept, detailing the preventative and corrective maintenance procedures completed. Procedures used for preventative maintenance should be detailed in the user manual. These procedures may be performed by system operators, or by specialised service personnel.

The procedures described under the quality control section should be used to check that the system continues to produce the output quality required of the system after the maintenance procedures have been completed. The results of the tests should be kept as part of the maintenance log.

The frequency of system maintenance will depend on the amount of usage of the system, but would typically be once per week.

These test results will serve to confirm at any later date that any poor quality images were not due to malfunction of the system.

If there is any deterioration in the output quality appropriate corrective maintenance should be done on the system.

Often, system maintenance hardware arid/or software can bypass the integrity checks used by the document. management system. The user manual should document procedures to control the use of these systems.

A record of system downtime, and details of action taken, should be recorded in the maintenance logs.

4.14 Security and protection

The system should operate within the guidelines provided in BS 7799: 1995, 'A Code of Practice for Information Security Management'. The procedures implemented should be described in the user manual.

When installed, the user should ensure that the system includes security controls appropriate to operational requirements. For example, user access controls should be provided.

To control access to the various levels of the document imaging system (e.g. manager, data input, retrieval), a secure password controlled access system should be implemented.

Where mixed media hierarchical storage systems are used, they should be assessed to ensure that they are used in a write-once mode only.

Removable media should be handled and stored in a manner recommended by the media manufacturers. Media should be clearly marked and stored in accordance with procedures in a secure manner.

Data file transfers, such as moving documents from one device to another should be controlled by the application software. It should not be possible to move documents or change index data without an entry in the audit trail.

All media should be kept in a secure area. At least one copy of the back-up should be kept off site.

Although the user facilities (document input and output) may be provided in a normal (unprotected) environment, the central part of the system (file servers, data storage, system software, etc.) should be installed in a secure area with restricted physical access.

Protection against virus infection should be installed.

The hardware should be protected from power failure by such devices as uninterruptable power supplies.

The hardware should be installed in a secure area, and access allowed only by authorised personnel.

All information about the status of documents, maintenance and quality control logs and audit trails should be kept in a secure manner, and be available for inspection by authorised external personnel (such as auditors) who have little or no familiarity with the particular document management system.

4.15 Backup and system recovery

Backup facilities on the system should allow for automatic backup and verification of all data files and associated information, including audit trails, at regular intervals. Procedures used in these systems should be documented in the user manual. There should be a record kept in the system audit trail of all backup activity, which should include details of any problems incurred during the procedure.

Backup procedures for system files should also be described in the user manual.

These procedures should include the implementation of off-site storage of these back-ups.

Back-ups will be acceptable if the structure of the data files held on the back-up is different to that of the original, provided that the structures are identified in the systems manual.

The restore processes should be described in the user manual.

In the event of a restore of data files and software from these backups, care should be taken to ensure the integrity of the system after the restore has been completed.

The recovery of the system from any type of failure should be described in the user manual.

The system audit trail should detail all data file recovery activities. This should include a note of any problems incurred during the procedure.

4.16 Use of bureau services

4.16.1 General

Many organisations will wish to send documents to a bureau that specialises in document management systems. Processing may include scanning and indexing of documents and/or transfer of data files to write-once optical disk. Where material is scanned, the output data file need not be transferred to optical disk by the bureau.

The procedures and recommendations in this section are intended to ensure that when work is done by a bureau, the resulting images stored by the client will be equally admissible legally, as if the work had been done wholly within the client's organisation, and to ensure that the client can prove compliance many years after the event, even if the bureau has ceased to trade.

The work done by the bureau may be undertaken on any site agreed between the client and the bureau. Generally the location of work would either be on the client's premises or at the bureau's premises. However, the Code covers situations where third party premises might be used.

Details of the procedures used in the transfer of documents and/or media from the client to the bureau, and from the bureau to the client, should be documented in the user manual.

The bureau should use procedures which adhere to this Code. The client should hold a copy of the bureau's User manual. This manual may need to be produced at the time of producing a document as evidence in a court of law.

A copy of the audit trails created during the bureau procedures should be supplied to the client, in such a form as to enable the client subsequently to interrogate it, to demonstrate the authenticity of a stored document.

A copy of the quality control and maintenance logs as applicable should also be supplied to the client.

4.16.2 Procedural considerations

Before agreeing a contract for the work with a bureau, the client should check the following:

♦ That the bureau can produce output of an acceptable quality level.

♦ That the bureau can process a sample of input material to produce output on the proposed media and in the proposed format and which can be successfully loaded on the client's target system.

♦ That a copy of the audit trails of the processing undertaken can be provided in an acceptable readable form.

- For scanning and indexing services, that the proposed indexing quality controls conform to the requirements of this Code, and that an acceptable level of index data accuracy can be provided.

- That the proposed location of the work is acceptable, and meets security criteria appropriate to the client's needs.

- That the proposed processing procedures will involve negligible risk of damage to the client's material.

- That where the material to be processed is unique or particularly valuable effective fire detection and prevention systems are implemented at the proposed production location.

- That where security of the material to be processed is important, the bureau will vouch for the trustworthiness of the intended operational staff. It is an advantage if all employees of the company sign a confidentiality agreement as part of their conditions of employment.

- Where documents sent for scanning form part of an active workflow system, the bureau should make arrangements to minimise disruption to the system due to the non-availability of documents.

4.16.3 Contract

A contract should be agreed between the client and the bureau, which should include the following:

- Where scanning and indexing services are to be provided, a statement that this Code will be adhered to.
- Description of production procedures (e.g. staple removal, document unbinding, photocopying prior to scanning).
- Description of the audit trails provided with the output, and the method of interrogation.
- Specification of the quality criteria to be met by the output generated by the bureau (For scanning applications this will include a Sample Set of test documents).
- Specifications of the format and media to be used by the bureau for delivery of the output data files, including reference to relevant standards.

4.16.4 Transportation of documents

The bureau should promptly check received material against the despatch schedule and advise the client of discrepancies as soon as practically possible.

Each shipment of material to/from the client and the bureau should be accompanied by a despatch schedule providing:

- Number of physical items in the shipment (e.g. file, boxes, tapes, disks);

- Identification list for shipped items;

- The client should check that material in transit is covered by insurance, either their own or that of the bureau.

All material being shipped should be adequately packed to avoid risk of damage in transit

The recipient should promptly check received material against the despatch schedule and advise the bureau of discrepancies as soon as practically possible.

4.17 Remote transmission of data files

This clause deals with data files transmitted from one site to another, as an integral part of a document management system.

If a document is input/scanned into one part of an organisations network, and is transmitted via the network to the storage system, the input/scanning and networking systems should be considered as if they are an integral part of the document management system.

If the transmission of the data file to the storage system is made using a file transfer program, then the transmission system should be designed in such a way to ensure that transmitted files and received files are unaltered by the transmission system. Upon successful receipt of the file, the storage system should treat the file as if it had been created within the document management system.

Private or public switched telephone network systems should be used only with care, to ensure that the data file being transferred is unaltered by the transmission processes. The following controls could be used to assist this checking process.

A message identification system should be used to provide mutual non-repudiation. The scheme chosen should include a message identifier and a transmission date and time stamp. Any message identification file should be stored on the document management system in association with its data file.

The message identification system should include a confirmation and/or an echo back, which should be used to confirm to the sender that correct receipt of the data file had occurred. This confirmation should include a message identifier and confirmation of receipt date and time stamp.

Once received by the document management system, the data file should be treated as if it had been created within the system.

File encryption systems will enhance the security and authenticity aspects of the file transmission system. The intended role and use of file encryption Systems should be agreed between sender and recipient prior to any transmission being initiated.

A carrier inserted date and time stamp should be transmitted and stored in association with the data file. A calling line identifier (CLI) should be treated as potential additional evidence only, since it is not always present, and the number presented may not necessarily be that of the sender

4.18 Voice files

4.18.1 General

This clause deals with various types of voice (audio) files. It covers speech use of audio as opposed to entertainment audio.

De facto and/or de jure standards should be used for voice digitisation and compression, wherever relevant.

4.18.2 Classes of voice files

There are three main classes of voice files that are classed as a documents:

- Annotated documents

These are voice documents; explanations etc., associated with other forms of data file (e.g. text, image). The associated files are normally stored as a compound document. Voice annotation may be as one of a sequence of data files or as an item pointed to from within a document. As such it may be stored on separate optical systems.

- Conversations

Where part or all of a contract is agreed verbally, with perhaps some confirmation documentation being generated, the voice file is the main document, possibly with an associated data file. The associated files are normally stored as a compound document. Note that one of the participants in a conversation can be a machine and telephone key tones may constitute a response and may need to be recorded.

- Voice instructions

Instructions can be left as a message on a voice recording machine, or may be delivered by voice mail systems.

4.18.3 Capture

The procedures used for the capture of voice messages should be described in the user manual.

Voice files may be generated internally i.e. from within the organisation or externally from a call across a public network. The files should be captured under the same conditions as that imposed for image files.

For all voice files, the system should use date and time stamping to ensure that no modifications are made to the original message. The recording system processes in this Code and the management controls in this Code should apply.

Where the voice file capture is not under the control of the document management system, the capture system should have at least as good control of file integrity as that imposed by this Code for other types of data flies.

4.18.4 Authenticatiotion

Where a source of the voice data file is relevant, this information should be included, as it may be required as evidence. The techniques should include:

- **Identity in the message by the caller.**
- **Additional identification (e.g. birthplace, PIN, 6ther unique identifiers).**
- **The calling line identity.**
- **Time and date stamping.**
- **Voice identification (where possible).**

Where authentication processes are separate from the message recording processes, the system manager should ensure no possibility exists of incorrectly matching the two processes. An example of such file matching would be for a document with an associated voice confirmation attached as authentication.

4.18.5 Data entry by voice recognition

Voice recognition is not totally accurate, and therefore should not be used for legally significant transactions without the implementation of stringent quality control procedures to ensure data accuracy. In particular, voice recognition procedures can be used in document indexing processes, provided quality control procedures can identify and correct all inaccuracies.

4.19 Document status change (work flow)

Some electronic document management systems incorporate a workflow capability which supports the change of a document status. For example, index information about a scanned supplier invoice may be sent by e-mail to a user who, by changing an index field, approves payment.

The workflow rule definitions, and the user procedures, should be included in the user manual.

Some work flow applications link documents by virtue of changes to the index information. The creation and destruction of these links should be recorded in the audit trail of each document affected.

Every change to the index information should automatically update the audit trail for the relevant documents.

Section 5. Enabling technologies

5.1 General

For a new system, the user should ensure that the system has been designed in accordance with the requirements of this Code. For systems already in operation, documents stored on the system prior to the introduction of this Code cannot be considered as conforming to the Code, unless controls which meet the requirements of the Code were in place from the time of storing the documents.

This section of the Code describes technologies, and how they are utilised and controlled.

- Systems description manual
- Access levels
- Systemintegrity Checks
- Audit logs
- Compound Documents
- Image processing
- Compression Techniques
- Forms overlays
- Environmental Considerations
- Data migration
- Document deletion and/or expungement

The user should check the conformance of the elements of the system to International Standards. This enables system auditors to check the performance and reliability of the system against these standards. Some elements of the system may not conform, due to the absence of published International Standards or the implementation of vendor specific. These elements should be specifically evaluated for conformance with this Code.

5.2 System Description Manual

A list of hardware and software elements which comprises the system should be included in the system description manual. This manual should also contain a description of the software elements of the system, and how they interact.

5.3 Access levels

Detail of all levels of access available on the system should be documented in the system description manual.

These levels are usually available as follows:

- System manager
- Document storage and Indexing.
- Document retrieval

The system should only be used for entry or amendment of documents by staff who are authorised to use the system, and such authority should only be granted after the member of staff has successfully proved his or her competence after adequate training. The system description manual should describe how these procedures are used for limiting accessing the system.

5.4 System integrity checks

Facilities should be provided within the system to ensure that the integrity of data file is preserved throughout the scanning system, and through the transfer of this data file to the storage media. Details of these procedures should be included in the system description manual.

A recommended approach is to utilise a checksum calculated immediately after the image has been captured (or after any necessary image processing), either from the uncompressed or the compressed image (or checksums from both), which should be stored with the image. This technique ensures that any errors in data file transfer between sub-systems may be detected automatically and with certainty.

This method on its own does cover the possibility of fraudulent manipulation of the image between the time of capture and the time of committal to the storage media. Such manipulation could be accompanied by the calculation of a new checksum if the checksum algorithm is known. To deal with this eventuality, other procedures are required. A simple method would be to write each checksum to the audit trait after calculation, since the checksum need not be more than a few bytes in size. The overhead involved in this process is relatively small. Alternatively, or in addition, after each batch of images has been captured, the aggregate checksum for all the images in the batch could be written to the audit trail.

Digital signatures, which provide powerful evidence concerning the authenticity and source of the document, may be used to facilitate the integrity of data files.

5.5 Audit logs

A provision for logging the principal activities on the system to the audit trail should be implemented. If this record is to be kept manually, then the user manual should describe how the records are kept. If the records are created automatically by the system, then the system description manual should describe how they are created, stored, accessed and maintained.

The contents of the audit trail are described in the relevant sections of this Code.

5.6 Compound documents

Where an image is captured as an entity on a scanner, and parts are electronically separated to be processed in different ways, these parts after processing should be stored on the same storage device, along with data identifying their respective locations within the original image, to allow accurate and unambiguous reconstitution of a facsimile of the complete image.

For example, this separation into parts may be done under operator control (i.e. on screen) or via software designed to separate out parts of an image containing photographic material from text. The photographic part may then be kept as a grey scale or full colour image, while the text part may be stored as a digital image. The original document may now be stored as a higher quality 'facsimile' than would have been the case if only a digital image of the whole page were stored.

The system should automatically determine the locations of the sub-images and should not allow access to this location data prior to its committal to the write-once media. Location data should also be written to the audit trail. Compound document identification and control should be detailed in the system description manual.

5.7 *Image processing*

To provide optimum image output, or improve recognition rates for an automated data capture process, post scanning processes can be performed. The effect on the image of each of these processes should be individually described in the system description manual.

The term 'post scanning processes' is used to describe various image enhancement techniques using hardware and/or software, that can singularly or independently have an effect on the presentation of image output and the size of the stored file. They can be installed on either a scanner workstation or a network server.

Some of the more common techniques are as follows and described in more detail below:

- Deskew
- Despeckle
- Black border removal
- Background Cleanup
- Noise removal
- Forms removal

These techniques are described in Annex B. They should be used with extreme care. For example, the despeckle process may remove decimal points, altering the value of numbers.

Any processing performed on the digitised image should be such as not to affect the integrity of the image as a true facsimile of the original. To check that any image processing does not affect the integrity of the scanned images, a sample set of documents should be scanned with the image processing active and hard copies prints of these images compared against the originals.

Processing performed on a grey scale image prior to conversion to a digital (black-and-white) image is generally acceptable, but the effect of such processing should be demonstrated by the system supplier

Where it is important that there be no loss of information in the scanned image other than that due to the scanning resolution, image processing procedures should not be used.

Speckle removal, which results in the elimination of single pixels or small groups of pixels from a digital image, to result in a subjectively cleaner image, should only be used with extreme care. Speckle removal cannot be relied upon just to remove noise from the image. There is high risk that information may be removed, e.g. parts of already broken characters, punctuation marks, parts of fine detail in drawings.

If speckle removal has been used, this fact should be recorded in association with the final image on the storage media.

Edge enhancement may be used in documents containing text and/or line art/drawings, to improve the representation of the text characters and line art; and so render this material more legible.

5.8 Compression techniques

Data flies may be compressed by the system prior or during storage, to reduce the capacity required of the storage system. The use of such techniques should be documented in the system description manual.

Compression could be of two types, namely lossy or lossless.

The system should provide adequate control facilities, preferably via automated means, to ensure that the requirements of legal admissibility can be met (e.g. inability to change scanned data, provision of detailed audit trail).

The type of compression used should be recorded on the storage media as part of the data file.

Lossy compression techniques should not be used for documents containing primarily text (including handwriting) or line drawings. By definition, lossy techniques mean that information is removed from the image during the compression process so that the decompressed image will not be the same as the original image. Parts of the text or drawings may be removed, being replaced by artificially generated data.

Lossy compression may be used for photographic or other continuous tone material, grey scale or coloured documents where it is evident that there be no visually detectable loss of information in the scanned image.

If lossy compression is used, a Sample Set of scanned images should be compared with the originals to check that there is no significant loss of information.

The compression ratio should be chosen such that all information which is required within the application context is present in the decompressed image. The maximum compression ratio acceptable may be determined via the sample set of originals.

The maximum acceptable compression ratio will typically vary between documents in the sample set, and a decision may need to be made whether to use different compression ratios for different documents or to use a single ratio for all documents. If the latter approach is adopted this will usually mean that the average image file size will be higher but the speed of scanning will also be higher because of reduced operator intervention.

Where it is important that there be no loss of information in the scanned image other than that due to the scanning resolution, lossy compression should not be used.

Examples of documents which should not be compressed using lossy techniques include radiographs, medical and engineering x-ray images.

The system should provide adequate control facilities, preferably via automated means, to ensure that the requirements of quality control (e.g. checking of image quality after scanning, with ability to re-scan if necessary; control over index data accuracy; control over data integrity) can be met.

5.9 Control of the 'write to storage' process

A description of the processes involved in the writing of flies to storage media should be included in the system description manual.

At the time of writing information to the storage media, it should:

- **record on the media the date and time and the location of the data file;**
- **verify that the data file has been written completely;**
- **track error corrections and/or report media errors;**
- **produce audit trails of the above processes;**
- **confirm that the indexing system has correctly identified the new documents.**

Appropriate additional manual procedures should be implemented where if necessary to support these processes.

5.10 Forms overlays

The processes used in the control of separate 'form' files should be described in the system description manual.

If images of forms are held separately from the data files to which they relate, they should be controlled as if they are part of the data file. If forms are modified (as part of normal business procedures), then data files should be kept of all versions of the form relevant to the data file being stored.

If 'forms removal' software is used, a record should be made automatically that the resulting image (the stripped form) has been the subject of forms removal and an identifier of the template used for that removal should also be kept. This information should be stored on write-once medium along with the resulting image. A copy of that template should also be recorded on the write-once medium. A

facsimile made by merging the template with the stripped form is not be a true facsimile of the original, although it may be sufficiently accurate for application use.

If it is required to retain true facsimiles of the original forms, consideration should be given to retaining the originals, making a microfilm copy (e.g. simultaneously with scanning), or retaining a complete image of the form also on write-once media, but stored off-line.

5.11 Environmental considerations

A description of the hardware manufacturer's recommendations for the operational environment should be included in the system description manual.

Handling and storage procedures should also be described in the system description manual.

Procedures for checking the condition of the media should be described in the system description manual. Media should be checked regularly in accordance with the media manufacturer's recommendations.

5.12 Data file migration

There should be provision for migrating data files held on a document storage system to new technology without loss of integrity, and with sufficient migration process documentation, such that the integrity of the documents can be established beyond any reasonable doubt at any time in the future.

Data files should be held in a format compatible with international standards, without encryption of the file names, to ensure ease of system audit and migration to new- systems.

5.13 Doc cement deletion arid/or expringement

It may be necessary to amend, delete or expunge specific documents from a document management system, due to a court order and/or to meet the requirements of the Data Protection Act. The requirements of the Data Protection Act are described in their document Data Protection Guidance for Users of Document Image Processing Systems'[3]. Paragraph numbers referred to in this section relate to this guidance document. Only document management systems which hold 'Personal Data' need registration under the Data Protection Act.

It is essential that document management systems have facilities to delete or expunge documents, as described in the Guide to the Data Protection Act (See paragraphs 25-29 of the Guide). This deletion or expungement can be achieved by the removal of index entries to the relevant documents.

It is essential that document management systems also have the facility to amend incorrect data, or remove irrelevant data, typically held in contravention of the Data Protection Act (See paragraphs 36-57 of the Guide). Such correction may be performed by deleting the original document file and substituting a document file containing corrected data, or by deleting selected parts of a document e.g. with the use of masks.

Section 6. Audit trails

6.1 General

A record should be kept of every significant activity on a document management system. These records form an audit trail. This audit trail should be created by the system. It is essential that this data is easily accessible by people not familiar with the system. The User manual should describe how the audit trail can be accessed and interpreted.

Audit trails should contain as much of the following information as is practicable:

- **Process**
- **Time and date of process initiation**
- **Amendments to index files**
- **Operator name**
- **Workstation reference**
- **Comments**

The time and date stamp should be sufficiently accurate that any subsequent investigation can determine the sequence of events.

Clause 6.2 to 6.7 list items which may be included in an audit trail.

6.2 Batch data

This data is to be recorded with every data file.

- unique batch ID
- date/time of status change
- operator ID at time of status change
- number of files within batch

Audit trail data should be copied periodically from the scanning system on to magnetic media and/or hard copy.

The audit trail should be subject to at least as good internal records management procedures as other 'critical records' of the organisation.

A secure back-up copy of the audit trail should be kept.

Audit trail information kept within the scanning system should not be modifiable, i.e. the audit trail itself should effectively be 'write once'.

The audit trail should be created automatically. It should record relevant data at each change of status of a document.

Data files which may be referenced in the audit trail are listed below. Specific data to be kept will be application and system dependent.

6.3 File data

- **unique ID for the file within batch**
- number of documents processed within file

6.4 Document data

- **unique ID for document within a file**
- number of pages (sheets) processed
- number of page images scanned
- number of pages transmitted to storage device

6.5 Batch status changes

These are primarily used for production control and failure recovery purposes, but could be useful to demonstrate that documents were correctly processed in accordance with the required procedures.

- received for scanning
- selected for preparation
- selected for pre-indexing
- pre-indexing completed
- selected for scanning
- scanning completed
- selected for quality control
- quality control completed
- selected for post-indexing
- post-indexing completed
- originals returned/sent for storage/destruction

6.6 Document status data

This data is temporary, and is used for production control purposes.

- total number of sheets
- total number of sides
- number of blank pages
- ready for scanning
- removed (coded with reason for removal)
- selected for quality control
  Quality control accepted
  Quality control rejected

6.7 Page data

This data gives details about the type of document being scanned.

- **original and photocopy**
- **photocopy flag (i.e. photocopied during preparation for scanning)**
- **photo reduction flag (i.e. photo reduced during preparation for scanning)**
- blank page indicator: indicates that this page is not to be scanned (i.e. is reverse of a single-sided sheet) or if scanned on a duplex scanner may be ignored

The term file' as used in this clause should be interpreted:

- as encompassing all similar types of device for holding documents, including, inter alia, envelopes, box files, ring or other type of binders;
- for microform media, as encompassing microfilm rolls and sets of fiche of film jackets contained within a single envelope.

The term page encompasses microform 'frame', and any other single image entity, such as a drawing or plan, map, photograph, transparency.

Annex A - Scanning specific documents

A.1 General

This section gives details of different types of documents, and the scanner characteristics needed to give acceptable results within the document management system

A.2 Tex4 typed and printed

In general, it is unlikely that a scanner resolution less than 200 dpi will be adequate to capture text to a subjectively acceptable quality level, in the sense that the eye would encounter no greater difficulty in reading from a hard copy print made (at the same or greater resolution) on a suitable printer.

- At lower resolutions, some characters may have missing detail, particularly if they contain thin elements, including serifs; typefaces under about 6 point on the original may not be captured very clearly.

With material containing particularly small type sizes (e.g. superscripts and subscripts), a resolution of 300 dpi or more may be necessary.

No decisions should be made regarding choice or resolution without conducting tests.

- Depending on the scanner, the quality. at 200 dpi may not be greatly inferior to the quality at 300 dpi.

It is important to bear in mind that a typical screen used for viewing document images has an effective resolution of about 100 dpi, or even less. This is typically adequate for much typed material but 'zooming' may be required with small sized print, and this requires that the scanning resolution should be substantially greater than the basic display resolution.

While post-scan image enhancement may improve the subjective image quality and legibility this should only be done after careful tests to ensure that the resulting image remains an effectively true facsimile of the original.

These tests should be done with the sample set of documents, and hard copies made of scanned images with and without image enhancement being used.

There should be no artefacts introduced into the enhanced image which are detectable under normal office lighting conditions with the unaided eye.

The results of these tests should be preserved with other records of the scanning processes.

A.3 Line drawings/art

For line art/drawings which form part of otherwise text - oriented documents, the scanning resolutions applicable to text are typically satisfactory for the drawings also. With printed material, where fine lines are used in the artwork, 200 dpi may be too low, but this can only be determined via tests on sample documents.

A.4 Hand-written material

With material where a modern pen, ball-point or pencil was used, 200 or 300 dpi will normally be adequate. But for older material where a steel-nibbed pen was used (e.g. to produce copperplate handwriting), the thinness of the upstrokes will often be such that 400 dpi will be the minimum resolution which will satisfactorily capture the text without significant components of these upstrokes being lost.

Handwriting (or hand drawing) using pencils can be faint, and difficult to reproduce. Care should be taken to ensure that image brightness and contrast are appropriate for these images.

A.5 Plans and drawings

For hand drawn architectural and engineering drawings, there may be finer lines present than would be the case with a typical full-sized CAD drawing, and although 200 dpi will usually be a satisfactory resolution, tests should be done to ensure that the finest detail is captured. It may prove necessary to use 300 dpi or even 400 dpi.

With CAD drawings, line thickness will vary depending on the size of the output. With the larger formats, e.g. AO or A1 (or US or Imperial equivalents) line thickness' tend not to be so fine as with some hand drawn material, and it is usually satisfactory to scan at 200 dpi. However, tests should still be done to check whether a higher resolution may be required. With smaller sized CAD drawings resolutions higher than 200 dpi may be required.

If the scanning is to be done from copies of the originals, and if these copies have been reduced from the originals (which is quite common), then a higher resolution than would otherwise have been satisfactory (i.e. if the original had been scanned) may be required: e.g. 300 dpi or 400 dpi compared to 200 dpi.

With drawings, dimensional accuracy may be important. Because of the large size of drawings, the paper or film may undergo dimensional change (due mainly to variations in moisture content). For use as working drawings it is often a requirement when scanning that dimensional inaccuracies are corrected: i.e. the scanned image may be post-processed to correct scale inaccuracies, skew, lack of orthogonality. Such corrections mean that the subsequent image is not a true facsimile of the original. Where the issue of legal acceptability may become an issue, it is recommended to preserve an uncorrected version of the scanned image as well as the corrected version.

A.6 Maps

With maps, a minimum resolution of 400 dpi will normally be required, but much higher resolutions (e.g. up to 1000 dpi) may be required with some material.

As with drawings, scanned images of maps are frequently corrected for scale inaccuracies and lack of orthogonality in the original after scanning.

Where coloured maps are being scanned, and the colour is to be preserved, the scanner should be capable of capturing individual colours with the required discrimination. While the number of colours subjectively present may he quite small, 8-bit. colour *(256* colours) may be inadequate and it may be necessary to scan with 24-bit colour in order to provide the required colour discrimination. Tests should be done to determine how many 'bits' of colour are required.

A.7 Halftone material

Where halftone material (black and white or colour separated) is present on a page along with text and/or line art, the objectives of the scanning should be addressed.

If the objective is to produce a scanned image which is comparable in quality to a 'normal' black-and-white photocopy, then a scanner which produces a digital image (i.e. 'black-and-white') will suffice. The resolution may have to be higher than that which would be acceptable for text only: 200 dpi will produce a rather poor quality image of the halftone material, 300 dpi will be significantly better, and 400 dpi is usually preferable.

If the halftone content has value in the application context, following the recommendations which apply to scanning text or line art may result in the capture of unacceptable quality images from the halftones.

Most scanners have different settings for scanning text/line art and scanning halftones. It is a general problem when scanning mixed text/line art and halftones with a 'black-and-White' scanner that the scanner settings which are optimal for text are far from optimal for the halftones, and vice versa. When set for 'text', the quality of the halftone images will generally be significantly worse than a photocopy; when set for 'halftone' or 'photographs', the text may appear rather blurred in the scanned image, to the extent that the image would not form a good facsimile of the original text.

If the halftone content has 'cosmetic' value only and does not contribute to the essential information content of the original, then the scanning should be done according to the recommendations which apply to text or line art material.

If the halftone is to be captured to a quality level comparable to a typical (good quality) photocopy, then there are two options. One option is to scan the document with the scanner settings 'normal', at a higher resolution than would be necessary for the text alone; 400 dpi minimum is recommended. The other is to scan the document twice, to create two images, one where the text/line ant is captured to satisfactory quality and the other where the halftone material is satisfactory. In the latter case a record should be kept that the two images involved different scanner settings (affecting the processing performed on the images).

If the halftone material is to be produced to a quality comparable to the original, then it should be processed according to the recommendations for photographs.

A.8 Photographs and radiographs (X-ray photographs)

With material. containing continuous tones (grey scale or colour), where the tonal information should be preserved, scanning should be done with a scanner capable of capturing the required number of grey levels and or colour. The number of levels which is appropriate should be determined by benchmark tests on the sample set of documents.

- Typically, the number of grey levels will be 16, 64 or *256* (i.e. 4, 6 or 8-bits per pixel). For good quality images from photographic material, 256 levels will normally be used. For very high quality images, and for X-rays, up to 1024 levels of grey (10-bits per pixel) may be necessary.

24-bit per pixel of colour information is usually quite adequate in most applications, but for very high quality images, up to 36-bits per pixel may be necessary. In many applications 15 or 16 bits of colour may be adequate; for source material containing only a small palette of colours, 256 levels may suffice. Only tests can determine how many colour levels are required.

- With continuous tone colour, most scanners capture 8-bits of colour information in three different regions of the colour spectrum: red, green, blue (RGB), resulting in 24-bits per pixel, or the ability to reproduce over 16 million colour variations.

- With only 8-bits of colour information (256 levels), there may be a noticeable 'blockiness' in the image if the original contains a broad range of colours.

The scanning resolution for coloured material will normally be similar to that for black-and-white material, particularly if there is text present on the original. Thus scanning may be done at 200-400 dpi, referred to the original photograph. If there is no text present on the original satisfactory images may be achieved at lower resolutions, down to television quality levels (about 350 lines per image frame); this would typically be satisfactory for mug shots' and similar applications.

To assess image quality, in general it is satisfactory to compare the screen images against the original. If there is likely to be use of high quality hard copy images then the comparison should be made between hard copies of the images, produced on an appropriate high quality colour printer, and the originals.

Care should be taken when comparing screen colours against an original that the colours were correctly balanced at the time of image capture, and that the display system has also been calibrated correctly. Otherwise the displayed colours may be significantly different from the colours on the original. The same requirement applies when comparing the original to hard copies of the captured image.

Where colour accuracy is important, a standard Colour Gamut test chart [4] should be scanned at the same time as the original (or batch of originals scanned at the same time), and the image of this chart stored along with the original.

A.9 Mixed mode documents

Mixed mode documents comprise more than one document type inside a single document (e.g. photograph, text). The documents described above containing halftone material are essentially of this type from a scanning perspective even though the original has been created in a single print operation. As described in the context of halftone material, the use of scanner settings optimised for one type of material can result in the loss of information in material of other types. As suggested for halftone material, one solution is to capture multiple images, with scanner settings (or even scanner type) selected to optimise the image quality for each material type.

One option is to use a scanning system which can scan mixed mode documents automatically, with automatic detection of each type of material and automatic optimisation of the settings for each type. These systems can also be set to select the most appropriate compression algorithm for each type of material. Benchmark testing should be done to ensure that the results are acceptable.

A.1O Documents with note sheets attached

Some documents may have note sheets attached - for example. self-adhesive notelets. Care should be taken when scanning these documents. If necessary, (for example when the notelet obscures information on the document) remove the note and scan it separately, along with a note to say to which document it was attached.

A.11 Microform documents

Microforms should be examined carefully prior to deciding upon the scanning approach.

Within multi-frame microfilm media (roll film, microfiche, microfiche jackets, multi-frame aperture cards), unless the inter-frame gap can be detected unambiguously automated frame detection should not be used.

If the gap is not detected multiple frames may be merged into one image. Depending on the physical characteristics of the scanning system it is possible that some part(s) of the digitised image may be lost.

With jacketed film, filmstrips may overlap. The processing procedures should ensure that such overlaps may be detected and corrected before scanning otherwise some page images will be missing or illegible, in whole or in part.

Where it is known that a rotary camera was used, this information should be recorded.

The images on the film may not have a one-to-one correspondence with the original documents. For example, when filming is done on a rotary camera it is possible for two pages to be fed at once, so that on the film part or all of an original page may be missing.

Annex B - Post-scanning processes

B.1 Document skew

Document skew is a term used to describe the phenomenon of poor document alignment (rotation) during the scanning processes. In its most pronounced form, images can appear on a viewing screen as crooked or slanted. A small angle of skew is likely to impact data capture processes and reduce data recognition rates.

Passing images through deskewing processes may correct this problem.

B.2 Speckle, noise and background marks

Random black marks (speckles) which appear on an image may have been generated during the scanning process, or are present on the original document. These speckles may be removed by systems involving special algorithms. These algorithms assume that small isolated clusters of pixels contain no information, and may be deleted.

In more sophisticated processes, pixel patterns are analysed for size and presentation in a filtering process.

B.3 Black border removal

When scanning documents of mixed sizes through certain scanner types (such as rotary scanners), black borders may be left around the edges of smaller documents. Black border removal entails the deletion of such large areas of black pixels.

B.4 Forms removal

The scanning of textual information on a pre-printed form is common when automated data capture processes such as OCR and OMR replace a large keyboarding operation. To increase the accuracy of the recognition rate, images can be passed through a post scanning process that will remove boxes, lines and pre-printed text.

System set-up procedures are used to establish the size and position of boxes and lines which should be removed.

Annex C - Optical character recognition

Optical character recognition (OCR) is the process of automatically entering printed text into a computer system without the need to key the data manually. An image file can be convened into ASCII text which then can be edited using a standard word processor. In principle, this is a two stage process, firstly acquiring an image with the use of a scanner, and secondly processing the image through OCR software, the results of which being an editable, searchable computer data file.

Some document management systems use OCR techniques to automate the indexing process.

To perform at its best, OCR software should be presented with as good a quality image as possible. That is an image where the characters are separated from each other and complete in outline. The image should be of a great enough resolution to provide sufficient detail for accurate recognition, but not so great that an excessive amount of computer power is required to handle it. As a rule of thumb, the size of the text measured in points multiplied by the resolution in dots per inch should equal 2400 or greater, i.e. 300 dpi 8pt type is clealy discernible, while to perform accurate OCR on 6pt type a resolution of 400 dpi is more appropriate.

Most scanners on the market today are more than capable of producing an image of usable quality for the purpose of OCR, but there are some points that should be considered in scanner selection.

If it is intended to scan and OCR large volumes of text, then a scanner with a multi-page sheet feeder might be appropriate.

If the OCR volume is low, then a hand scanner might be sufficient, but remember it takes a steady, practised hand to give an even image and an A4 page will take 2 or 3 passes which will require 'stitching' together before recognition can take place.

There are many smaller sheet feed scanners appearing on the market which are ideally suited to OCR use. They provide very acceptable image quality, and are small in size. Often they are equipped with a 10 or 20 page sheet feeder for low to medium volume jobs.

Colour scanners have little to offer the OCR process other than if they are capable of selecting the colour of the lamp (or the filter over the image sensor). If this is possible then it can be an advantage in situations with text on coloured backgrounds. Scanning with a red lamp or filter will 'drop out' or reduce a red or pink background to white, giving a better differential between the text and its background.

No matter how good the image on which the OCR process is to be performed, the results are unlikely to be 100% accurate. Therefore it is only appropriate to use this technology for capturing index data in applications with use multiple index keys. The use of this technology is becoming common as a data capture method for 'free text retrieval' system.

The text file that results from the OCR process output will be considered as a poor copy of the original in terms of its weight as legal evidence. Therefore this technology is appropriate only as a method of facilitating the searching for documents.

Where OCR technology is used to assist indexing and searching, it is essential that the TIFF image input file is the entity which is retrieved and used as 'the document'. Any text file produced is likely to be of reduced evidential value compared with that of the original image file.

Annex D - List of references

Normative references
BSI standards publications

BRITISH STANDARDS INSTITUTION, London

BSI Publications are available from Customer Services, Sales Department, 389 Chiswick High Road, London W4 4AL.   Tel: 01 81 9967000; Fax: 0181 996 7001

BS 7768: 1994
Management of optical disk (WORM) systems for the recording of documents that may be required as evidence

BS 7799 :1995
Code of Practice for Information Security Management

Informative references
BSI standards publications

BRITISH STANDARDS INSTITUTION, London

BS 6001 Sampling Procedures and Tables for Inspection by Attributes

Part 1:1991
Specification for sampling plans indexed by acceptance quality level (AQL) for lot-by-lot inspection

Part 2:1993
Specification for sampling plans indexed by limiting quality (LQ) for isolated tot inspection

Part 3:1993
Specification for skip-lot procedures

Part 4:1994
Specification for sequential sampling plans

BS EN ISO 9000
Quality management and quality assurance standards

Other publications

[1] IMAGE AND DOCUMENT MANAGEMENT ASSOCIATION (IDMA), *Principles of Good Practice for Information Management,* IDMA, London, 1995. Available from IDMA, The Department of Information Systems, The London School of Economics, Houghton Street, London WC2A 2AE.

[2] GREAT BRITAIN. Civil Evidence Act 1995. London: HMSO Available from HMSO, 49 High Holborn. London WC1 for personal callers or by post from HMSO, P0 Box 276, London SW8 SDT.

[3] DATA PROTECTION REGISTRAR, *Data Protection Guidance for Users of Document Image Processing Systems*, 1995. Available from Data Protection Registrar, Wycliffe House, Water Lane, Wilmslow, Cheshire SK9 SAF.

[41 ROCHESTER INSTITUTE OF TECHNOLOGY, *Process Ink Gamut Chart,* Available from Rochester Institute of Technology, T & F Center, One Lomb. Memorial Drive, Rochester, New York 14623, USA.