TECHNICAL REPORT

ISO/TR 18492

First edition
2005-10-01

Long-term preservation of electronic document-based information

Conservation à long terme d'information document basée électronique

Reference number
ISO/TR 18492:2005(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In exceptional circumstances, when a technical committee has collected data of a different kind from that which is normally published as an International Standard ("state of the art", for example), it may decide by a simple majority vote of its participating members to publish a Technical Report. A Technical Report is entirely informative in nature and does not have to be reviewed until the data it provides are considered to be no longer valid or useful.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TR 18492 was prepared by Technical Committee ISO/TC 171, *Document management applications*, Subcommittee SC 3, *General issues*.

Introduction

Ensuring the long-term preservation of authentic electronic document-based information is a well-documented and identified problem within many fields of expertise, including archival science, document management, e-commerce, e-governance and technology development. As an additional problem, individuals and organizations charged with the responsibility for ensuring long-term access to authentic electronic document-based information have employed a diversity of strategies designed to achieve this goal.

Although there is a clear need to address the problem of long-term access to authentic electronic document-based information, there is a current lack of harmonized international guidance on these issues. This has led to diverse and, sometimes, incompatible approaches that can give rise to potentially mission-critical problems, regarding the accessibility and/or authenticity of the electronic document-based information being retained.

Acknowledging the generic technological obsolescence problem of computer hardware and software as well as the limited life of digital storage media, this Technical Report provides guidance to storage repositories in providing access to and maintaining authentic electronic document-based information that has been retained for future reference.

The purpose of this Technical Report is to provide a clear framework for strategy development and best practices that can be applied to a broad range of public and private sector electronic document-based information to ensure its long-term accessibility and authenticity.

Long-term preservation of electronic document-based information

1 Scope

This Technical Report provides practical methodological guidance for the long-term preservation and retrieval of authentic electronic document-based information, when the retention period exceeds the expected life of the technology (hardware and software) used to create and maintain the information.

It takes into account the role of technology neutral information technology standards in supporting long-term access.

This guidance also acknowledges that ensuring the long-term preservation and retrieval of authentic electronic document-based information should involve IT specialists, document managers, records managers and archivists.

It does not cover processes for the creation, capture and classification of authentic electronic document-based information.

This Technical Report applies to all forms of information generated by information systems and saved as evidence of business transactions and activities.

NOTE Electronic document-based information constitutes the "business memory" of daily business actions or events and enables entities to later review, analyse or document these actions and events. As such, this electronic document-based information is evidence of business transactions that enable entities to support current and future management decisions, satisfy customers, achieve regulatory compliance and protect against adverse litigation. To achieve this goal, this electronic document-based information should be retained and appropriately preserved.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 12651:1999, *Electronic imaging — Vocabulary*

ISO 15489-1, *Information and documentation — Records management — Part 1: General*

ISO/TR 15489-2, *Information and documentation — Records management — Part 2: Guidelines*

ISO/TS 23081-1, *Information and documentation — Records management processes — Metadata for records — Part 1: Principles*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 12651, ISO 15489-1 and ISO/TR 15489-2 and the following apply.

3.1
authentic electronic document-based information
electronic document-based information the accuracy, reliability and integrity of which are maintained over time

3.2
document-based information
substantive information that can be treated as a unit (e.g. an image, text, spreadsheet, database views)

NOTE Document-based information is inclusive of, but not necessarily limited to: text, images, tabular data (e.g. a spreadsheet), or any combination thereof.

3.3
document-based information content
substantive content contained in document-based information

3.4
document-based information context
information about the circumstances of electronic document-based information creation, control, use, storage and management, and information about its relationship to other similar material

3.5
document-based information structure
logical and physical attributes of document-based information

NOTE Logical attributes consist of the logical order, e.g. a hierarchy with identifiable subparts, whereas physical attributes comprise elements, e.g. type font, spacing.

3.6
electronic archiving
storage of electronic information in an independent physical or logical space where the information is protected from loss, alteration and deterioration

NOTE The information may be used as reliable evidence in the future if it has been protected in this manner.

3.7
long-term preservation
period of time that electronic document-based information is maintained as accessible and authentic evidence

NOTE This period of time can range between a few years to hundreds of years, depending upon the needs and requirements of the organization. For some organizations, this period of time would be determined by regulatory compliance, legal requirements and business needs. For other oranizations, such as archival repositories holding public records, the period of time required to retain electronic document-based information is usually thought to be hundreds of years.

3.8
metadata
data describing the content (including indexing terms for retrieval), context and structure of electronic document-based information and their management over time

3.9
migration
process of transferring electronic document-based information from one software/hardware environment or storage medium to another environment or storage medium with little or no alteration of structure and no alteration in content and context

3.10
storage repository
storage repository organization or entity charged with the storage and maintenance of authentic electronic document-based information

NOTE It is recognized that this definition is different from technical definitions of “storage repositories”.

3.11
technological obsolescence
displacement of an established technical solution in a marketplace as a result of major technological developments or improvements

4 Symbols and abbreviated terms

ASCII	American Standard Code for Information Interchange
CRC	Cyclical Redundancy Code
HTML	Hyper Text Markup Language
JPEG	Joint Photographic Engineers Group
OCR	Optical Character Recognition
PDF/A-1	Portable Document Format — Archive
SHA-1	Standard Hash Algorithm 1
TIFF	Tagged Image File Format
WORM	Write Once Read Many (times)
XML	Extensible Markup Language

5 Long-term preservation

5.1 General

Increasingly, the proliferation of computer technologies that support the creation, use, storage and maintenance of information, results in private and public sector organizations relying on electronic document-based information as the official evidence of their business activities. Consequently, organizations increasingly face the challenge of ensuring the long-term accessibility of authentic electronic information that was created within reliable and trustworthy information systems and stored on electronic media that might be subject to technological obsolescence that if left uncorrected will make the document-based information irretrievable. The importance of this problem is compounded by the fact that organizations are increasingly conducting activities and transactions where no paper evidence exists.

It is essential, therefore, that organizations develop and apply a well-defined strategy for providing long-term preservation and retrieval of authentic electronic document-based information. Subclause 5.2 defines the elements of such a strategy.

5.2 Goals of a long-term preservation strategy

5.2.1 General

This subclause identifies six key issues that storage repositories should consider when they are developing a long-term preservation strategy.

5.2.2 Readable electronic document-based information

A long-term preservation strategy should ensure that electronic document-based information remains readable into the future. To achieve this, the bit stream comprising electronic document-based information should be accessible on the computer system or device that:

— initially created it or

— currently stores it or

— currently accesses it or

— will be used to store the electronic information in the future.

These four processablity options are predicated on the fact that electronic document-based information stored on digital storage media can become unreadable. There are two primary ways in which this can occur.

One is the result of exposure to hostile storage conditions. All of the media currently used for storing electronic document-based information share a common vulnerability to poor environmental conditions, e.g. fluctuations in temperature and humidity. These adverse conditions either damage the media or accelerate the ageing process. Different types of digital storage media require different levels of controlled storage environment to ensure maximum longevity. Some storage technologies are prone to data corruption through magnetic field interference, dust and environmental contaminants (magnetic storage media), while others (optical storage media) are not as prone to these outside factors and less susceptible to media damage outside tightly controlled storage environments. Regardless of which storage technology is in use, it is important to recognize that all forms of storage media can deteriorate and/or degrade through environmental changes.

The second is that non-readability may occur through media obsolescence, which occurs when a storage device (e.g. a tape or disk) is physically incompatible with the available computer hardware (e.g. a tape or disk drive) and therefore cannot be read. Based on past trends, media obsolescence in the future seems inevitable because advances in storage technology continually introduce changes in the way the electronic document-based information is physically stored (e.g. changes in recording technology, changes in disk drive hardware/software interfaces), the form factor of the storage media and in the way the underlying bit stream of document-based information is physically represented (e.g. error correction codes) or the form factor of the storage media. Consequently, over time, older storage media will become incompatible with subsequently used media.

A long-term preservation strategy should specifically address media obsolescence by establishing procedures for periodically transferring document-based information from older to newer media.

NOTE Data readability is important along with data formatting. Ensuring that the data are formatted in a fashion (i.e. technology neutral formats) that enables users in the future to process the data, should be taken into consideration.

5.2.3 Intelligible electronic document-based information

A long-term preservation strategy should provide intelligible electronic document-based information. Digital information is only intelligible to a computer if the computer also has access to information describing how to interpret the underlying bit stream. The intelligibility of electronic document-based information, therefore, is a function of information about what the bit stream in fact represents and the processing software's capacity to take appropriate action based on this information.

EXAMPLE The binary code (1s and 0s) comprising a digital Tagged Image File Formatted (TIFF) image carries no intelligibility in its own right. Rather, the image's file header, which contains information such as byte order and the compression algorithm used, enables a computer (through a combination of its operating system and image software) to display and print the image. Similarly, a word processing document carries metadata that makes it intelligible to word processing software.

5.2.4 Identifiable electronic document-based information

A long-term preservation strategy should provide identifiable document-based information. Identifiable document-based information should be organized, classified and described in such a way that it is possible for users and information systems to distinguish between information objects based upon a unique attribute such as name or ID number. Aggregating electronic document-based information into categories based upon shared attributes can facilitate searching and retrieval. Failure to provide such identification can severely limit searching and retrieval.

5.2.5 Retrievable document-based information

A long-term preservation strategy should provide retrievable document-based information, meaning that discrete information objects (or parts of them) can be retrieved and displayed. Retrievability is typically software-dependent in that it requires keys or pointers that link the logical structure of information objects (e.g. data fields or text strings) to their physical storage location.

Generally, this linkage is found in a database record, file system directory structure, file allocation table, header or label that includes the information required to locate the beginning of an object, to indicate the number of bytes of each component or data element and to establish its physical location on the storage medium.

The interpretation of the logical structure of document-based information is a function of an operating system or device driver in conjunction with a particular application system developed to store, manage and access digital information. The retrievability of information objects is therefore inextricably linked to a device driver, software application, file system or operating system.

Newer generations of file formats that support the readability of older file formats help ensure the ability to retrieve electronic document-based information. Backward compatibility however, can be limited because many software vendors support only certain file formats, while others support all versions of various data formats. An example of this would be support for TIFF, JPEG or HTML formatted data, which include backward compatibility.

5.2.6 Understandable document-based information

A long-term preservation strategy should ensure that document-based information is understandable. In order for electronic document-based information to be understandable, it should convey information to both computers and humans. However, the meaning of discrete document-based information is not determined solely by its content. Rather, meaning is derived from the context of both its creation and its use (i.e. metadata). As such, storage repositories should be aware that ensuring the understandability of electronic document-based information differs sharply from ensuring the understandability of paper documentation. Unlike paper documentation, where their physical characteristics typically convey the context of its creation and use, the context of creating and using electronic document-based information is usually linked logically rather than physically.

EXAMPLE A series of paper documents regarding a particular transaction may be stapled together or placed in a file folder, whereas electronic document-based information of a similar transaction may exist on multiple media in multiple locations and therefore should be electronically tied together. These logical linkages can include identification of both the business process that led to the transaction as well as the participants in the transaction.

The context of creation and use also involves relationships among other document-based information that has been be captured in a variety of ways, including a reference code in a document profile to the other material dealing with the same issue, or a classification code that links each instance of document-based information relating to the same transaction.

Successful retrieval of electronically stored document-based information therefore depends in part upon preservation of these logical linkages regardless of the length of time they are retained.

5.2.7 Authentic electronic document-based information

5.2.7.1 General

A key goal of a long-term preservation strategy is to ensure the protection of authentic document-based information. Authentic electronic document-based information is what it purports to be, i.e. reliable information that over time has not been altered, changed or otherwise corrupted. Organizations seeking to provide long-term access to document-based information that is authentic should consider three critical aspects in their strategy:

a) transfer and custody;

b) the storage environment;

c) access and protection.

5.2.7.2 Document-based information transfer and custody

It is difficult to protect electronic document-based information from alteration, so long as it remains in a production environment and is not stored on non-alterable, write-once media. Accordingly, a long-term preservation strategy should provide for the transfer of document-based information from production environments and from the originators and recipients to a storage system or storage repository, i.e. an operationally independent third-party charged with maintaining document-based information according to documented policies and practices.

5.2.7.3 Storage environment

A long-term preservation strategy should specify a stable storage environment for media containing electronic document-based information because hostile or improperly controlled environments put the information at risk.

5.2.7.4 Document-based information access and protection

A long-term preservation strategy should provide mechanisms to restrict access to electronic document-based information and protect it from deliberate or accidental alteration and corruption.

Electronic document-based information stored on rewritable media can be altered without leaving any physical evidence. Electronic document-based information is also vulnerable to accidental corruption during a transfer between media and information systems. As such, organizations seeking to ensure the authenticity of electronic document-based information over time should establish appropriate policy, practice and technology-based controls. Examples of common technology-based controls include:

— use of WORM (i.e. non-rewritable) magnetic or optical media;

— secure client-server architectures that can be used to block direct access to electronic document-based information, with the net effect of providing “read-only” access;

— Cyclical Redundancy Check code values (CRCs) commonly used as a technique for establishing the reliability of electronic transmissions and are therefore, particularly useful for verifying that no changes have been made to the electronic document-based information since being initially stored;

— one-way hash functions (e.g. SHA-1) employing an algorithm that can compress electronic document-based information into a fixed-length number of bits that effectively becomes a unique "fingerprint" of the electronic document-based information, and can subsequently be used to demonstrate it has not been altered.

6 Elements of a long-term preservation strategy

6.1 General

Maintaining accurate, reliable and trustworthy electronic document-based information means ensuring the following.

— It can be read and correctly interpreted by a computer application.

— It can be rendered in a format understandable to humans.

— It has the logical and physical structure, substantive content and context that were apparent at the time of creation or receipt.

Limited electronic media durability and inevitable technology obsolescence will force storage repositories, charged with providing long-term preservation of authentic and processable electronic document-based information, to make critical choices regarding long-term access. To deal with the challenges of media durability and technology obsolescence, storage repositories will find it necessary to employ diverse strategies and tools. These strategies and tools can be conceptually divided into three primary activities that collectively form the foundation of any long-term preservation strategy.

a) First, storage repositories should undertake media renewal (see 6.2) to address media durability.

b) Second, where automated tools exist, document-based information migration (see 6.4) is a viable option to address technology obsolescence by transferring document-based information from one technology platform to another.

c) Third, when digital information and images are stored within legacy information systems where no automated migration tools exist, a more robust approach may be required. The emulation of legacy information systems within current technology environments may be required. Although this course of action has a conceptual appeal, up to this point it has encountered operational resistance for the purpose of long-term access to authentic electronic document-based information. Therefore, emulation is not addressed further in this document.

6.2 Media renewal

6.2.1 General

Limited media durability and technology obsolescence suggest that periodic media renewal is both inevitable and a base-line requirement for ensuring long-term preservation of authentic and processable electronic documentation by keeping the original bit stream “alive”. Media renewal requires that electronic document-based information be either reformatted or copied as detailed in 6.2.2 and 6.2.3

6.2.2 Reformatting electronic document-based information

6.2.2.1 General

When document-based information is reformatted, its underlying bit stream changes because it is moved to a different physical carrier (e.g. from a media type containing 18 storage tracks to one containing 36 storage tracks) or the character code is transformed (e.g. from 7 to 8 bit ASCII), but there is no alteration in its physical representation or substantive content. Reformatting occurs independently of the software application that created the document-based information.

6.2.2.2 Causes for reformatting

There are three instances when organizations should consider reformatting electronic document-based information, as follows:

a) *Reformatting upon transfer*: electronic document-based information should be reformatted to a standard encoding representation and storage medium when it is transferred to a storage repository.

b) *Reformatting upon upgrade*: reformatting is justified when the storage repository upgrades its equipment and replaces existing storage devices with newer ones.

c) *Scheduled reformatting*: reformatting should be scheduled to coincide with the projected life expectancy of the media in use and projected life expectancy of the device or drive required to process the storage media.

6.2.2.3 Storage media for reformatting

Storage repositories should give careful consideration to the selection of storage media, especially when reformatting electronic document-based information. At the broadest level, organizations should choose between magnetic and optical technology. Among the issues that should be taken into account are:

— high storage capacity;

— high data transfer rate;

— minimum projected twenty-year life expectancy;

— established and stable market place presence;

— affordability;

— suitability.

A high storage capacity and a data transfer rate are critical because ultimately they drive the time required to transfer electronic–based information during media reformatting and copying, which is likely to become an issue as the volume of electronic-based information held by storage repositories is measured in terabytes and petabytes.

6.2.2.4 Reformatting and authenticity

The authenticity of electronic document-based information can be challenged after it has been reformatted, particularly if there are several instances of reformatting. To provide a satisfactory level of authenticity, storage systems and storage repositories should have a written quality control policy in place that mandates verification of the accuracy of all reformatted document-based information.

The procedures for the implementation of this policy should include thorough and complete documentation of all the steps followed in reformatting, including:

— identification of the individual(s) who actually executed the process;

— the date it occurred;

— the format of the data;

— comparison of CRC or hash digest values generated before and after reformatting to confirm that no changes occurred;

— visual comparison of several reformatted instances of document-based information with their counterparts in the old format.

Best practices should identify inaccuracies or unrecoverable errors and maintain follow-up document-based information of how they were addressed. The physical location (e.g. block or sector track) of any unrecoverable errors should be identified. In addition, a third party should review these actions in order to determine that they were carried out in accordance with established procedures. Finally, this documentation should be clearly identified with specific links to the document-based information, and treated as metadata that merits the same care as that of the document-based information itself.

6.2.2.5 Reformatting security

Storage repositories should protect electronic document-based information from alteration or loss during reformatting. Electronic storage media are vulnerable to human intrusions and catastrophic failure or natural disaster. As such, storage repositories should employ the following measures to minimize security risks.

— A “firewall” or one way link (e.g. “air gap”) should be installed that only permits read-only access, and only to authorized individuals.

— Electronic storage media should be housed in a locked, secured area or vault with controlled access.

— A backup copy of storage media should be created and stored in a separate location from the original.

— Two different types of storage media should be used for original and back-up copies in order to minimize the risk of unexpected technology obsolescence.

6.2.3 Copying electronic document-based information

6.2.3.1 General

The objective of copying electronic document-based information is to maintain its authenticity and processablity by transferring it from old storage media to new storage media with the same format specifications and without any loss in structure, content or context. The underlying bit stream of the electronic document-based information remains unchanged when it is copied to the target storage media.

6.2.3.2 Reasons for copying

There are three instances when storage repositories should consider copying electronic document-based information, as follows.

— *Copying upon transfer*: electronic document-based information should be copied when it is transferred to a storage repository that uses storage media of the same format specification as the storage media used before transfer.

— *Copying upon media errors*: electronic document-based information should be copied when the annual sample that is checked for readability problems discloses either a high number of temporary or uncorrectable read errors, but no media or device upgrades are necessary.

— *Scheduled copying*: electronic document-based information should be copied when the storage media is ageing, but no media or device upgrades are necessary because current versions are still widely supported and meet organizational performance requirements. Organizations should establish a fixed period of time (e.g. one-half the projected life expectancy of the media) to initiate the copying of electronic document-based information on to a new version of the acceptable storage media.

6.2.3.3 Copying authenticity

Although the bit stream of document-based information does not change when it is copied, there still remains the potential for corruption during the process. To provide a satisfactory level of authenticity, storage repositories should have a written quality control policy in place that mandates verification of the accuracy of all copied document-based information.

The procedures for the implementation of this policy should include thorough and complete documentation of all the steps followed in copying, including:

— identification of the individual(s) who actually executed the process;

— the date it occurred;

— the format of the data;

— the number of bits/bytes involved;

— comparison of CRC or hash digest values generated before and after copying to confirm that no changes occurred;

— visual comparison of several copied document-based pieces of information with their counterparts on the old media.

Best practices should identify inaccuracies or unrecoverable errors and maintain follow-up information of how they were addressed. Specifically, the physical location (e.g. block or sector track) of any unrecoverable errors should be identified. In addition, a third party should review these actions in order to determine that they were carried out in accordance with established procedures. Finally, the information detailing any identified problems should be clearly identified, have specific links to the document-based information and be treated as metadata that merit the same preservation care as that accorded to document-based information.

6.2.3.4 Copying security

Storage repositories should protect electronic documentation from alteration or loss during copying. Electronic storage media are vulnerable to human intrusions and catastrophic failure or natural disaster. Digital storage repositories should employ the following measures to minimize security risks.

— A “firewall” or one-way link (e.g. “air gap”) that only permits read-only access and only to authorized individuals.

— Electronic storage media housed in a locked, secure area or vault with controlled access.

— CRCs or hash values of electronic document-based information created before they are copied, and after they are copied, to verify that no changes occurred.

— A backup copy of storage media created and stored in a separate location from the original.

— Two different types of storage media used for original and backup copies in order to minimize the risk of unexpected technology obsolescence.

6.3 Metadata

6.3.1 General

Metadata (data about data) consist of information about the context, processing and use that supports the identification, retrieval and preservation of authentic electronic-document based information.

In some instances, some software applications can automatically create metadata such as file size, file format, data, hash digest comparison and other similar attributes (e.g. electronic document-based information properties and the like). In other instances, manual entry of other metadata such as classification, retention period, records series and key words, among others, may be necessary. These data and electronic document-based information are retrievable. It is likely that as organizations move toward the implementation of enterprise content management systems, metadata elements that can support preservation strategy will be far richer then those currently in use. In addition, it is likely that they will be automatically generated so manual entry will not be necessary. Accordingly, storage repositories should ensure that the tools supporting the capture and use of metadata are sufficiently flexible and scaleable to accommodate richer metadata elements as they become available.

6.3.2 Interoperable metadata

In the future, metadata residing on enterprise content management systems will be interoperable. Therefore, organizations designing the capture and use of metadata that in the future will be used in an interoperable environment should take into consideration ISO/TS 23081-1.

6.4 Migrating electronic document-based information

6.4.1 General

A long-term access strategy should include provisions for the migration of electronic document-based information.

Storage repositories with a mandate to acquire and preserve authentic electronic document-based information and to ensure access to it over time, face four challenges.

a) For the foreseeable future, organizations and individuals will continue to use a wide variety of software packages and storage formats for creating and using electronic document-based information. It will be very difficult for storage repositories that acquire this document-based information to have access to, or support for, all the packages and formats.

b) Some electronic document-based information is likely to be software-dependent and therefore usable only within a specific software environment.

c) Operating systems and software applications will inevitably be displaced by newer and faster offerings with increased functionality, meaning that storage repositories will periodically have to transfer electronic document-based information from the current software environment to a newer one.

d) Some electronic document-based information is likely to be retrievable only in a legacy information system that lacks automatic migration tools.

Electronic document-based information migration can successfully address all four of these challenges. Accordingly, storage repositories will have to support the migration of authentic electronic document-based information from one application environment to a new application environment with little or no loss in structure and no loss of content or context.

6.4.2 Software dependence

A long-term preservation strategy should address the issue of software dependence. When electronic document-based information can only be used within a specific software application, providing long-term access to this document-based information may be difficult, particularly if a vendor discontinues support or does not provide continuity in newer versions of the software. In many instances, it is possible to eliminate software dependence by sacrificing some loss of structure; e.g. textual document-based information in a native word processing application can be migrated to straight text (i.e. plain ASCII text) through automatic removal of embedded word processing instructions or code that control some aspects of the physical representation such as bold type and footnotes.

While such actions will reduce software dependence, storage repositories should carefully consider the impact on the authenticity of such migrated document-based information. Such document-based information can no longer be considered to be imitative copies because it does not replicate the structure of the original document-based information. Rather, the resulting document-based information should be considered to be “new” document-based information for which authenticity should be re-established through documentation of the actions taken, and validation that the substantive content of the document-based information has not been altered.

An alternative to migrating textual electronic document-based information to straight text is to print it on paper or microfilm, which would preserve authenticity at the expense of processablity. This approach is particularly appropriate with page analog electronic document-based information whose processablity could conceivably be restored in the future through the use of Optical Character Recognition (OCR).

Hierarchical and relational database tables also can be migrated to a flat table structure to minimize vendor specific software dependence, in which case the identification of primary and foreign keys in each table should be retained while the relational links are deleted. Whenever this is done, metadata should be created that identify whether these relationships were one to one, one to many, many to one, or many to many so that the links could be re-established in the future.

6.4.3 Software upgrades and new software installation

As software upgrades and installation of new software are inevitable for storage repositories committed to providing long-term access to authentic document-based information, a long-term access strategy should provide polices and procedures for this eventuality.

When software is upgraded (e.g. from Version 1 to Version 2) and the vendor provides backward compatibility between the upgrade and the old software, document-based information should be automatically moved, along with the underlying physical representation scheme, substantive content and context, to the new environment.

When new software replaces existing software, either as a stand-alone application or as part of a general information system upgrade, document-based information should be migrated using the export feature of the old system and the import feature of the new system. In addition, some environments may support migration through import/export gateways designed for specific proprietary formats (e.g. from one brand of word processor to another).

6.4.4 Migration to standard formats

Storage repositories should consider migrating electronic document-based information from the wide variety of formats used by creators or recipients to a smaller number of “standardized” formats upon their transfer to the custody of the repository. “Standardized” formats could be a consensus on formats that are widely used and are likely to cover a majority of a particular class of electronic document-based information. Proprietary file formats should be avoided. Among the technology neutral formats that merit consideration are PDF/A-1, XML, TIFF and JPEG.

6.4.5 Migrating legacy information system electronic document-based information

6.4.5.1 General

A long-term access strategy for authentic and processable electronic document-based information should require migration when there is neither backwards compatibility nor the existence of an export/import gateway between the old, or legacy, system that contains the documentation and the target information system.

In the future, the need for migration of legacy information system electronic document-based information may be lessened due to the broader development of systems that support vendor technology neutral architectures and formats. In the meantime, however, storage repositories will have to migrate electronic document-based information embedded in legacy information systems in order to meet their obligations.

Some loss of information during repeated migration cycles is inevitable due to the fundamental incompatibilities that exist between several generations of older and newer systems. Accordingly, storage repositories, rather than attempting to achieve no loss of information whatsoever, should consider developing migration policies and quality control procedures regarding how to reduce the degradation of information during migration. One important procedure is the documentation of loss during migration and the results of Quality Control activities. Wherever possible, this documentation should be retained with the media.

6.4.5.2 Migration steps

6.4.5.2.1 General

Storage repositories should implement a ten-part incremental approach to migration. As the circumstances of each migration can vary considerably, the ten steps presented below should not be viewed as a specific migration plan that necessarily applies in all circumstances.

6.4.5.2.2 Analyse the legacy information system (Part 1)

Storage repositories should analyse the legacy information system in order to understand its functionalities and the document-based information contained within it. This should include:

— the rationale for its functionalities;

— how metadata are captured and their relationship to the document-based information;

— the relationships amongst the document-based information.

The information product produced in this phase should be a specification that will be used in a “forward engineering” of the functionalities, metadata and document-based information to the new system.

6.4.5.2.3 Decompose the legacy information system structure (Part 2)

Storage repositories should decompose the legacy information architecture so that its interfaces, applications, and database services can be treated as distinct components — with awareness that this may not be possible for all information systems, as explained below.

— A legacy system is decomposable if the system and user interfaces, the application modules, the database service and the database itself are separate and independent components.

— A legacy system is semi-decomposable if the interfaces and database are independent, but the application and database services form a single module.

— A legacy system is non-decomposable if the interfaces, applications and database services are linked in one module.

In any case, any external dependencies in system architecture should be eliminated in preparation for migration.

6.4.5.2.4 Design the target interfaces (Part 3)

The target interfaces should provide a connection to the legacy interfaces.

6.4.5.2.5 Design the target applications (Part 4)

The target applications should provide a connection to the legacy applications.

6.4.5.2.6 Design the target databases (Part 5)

The target database should provide a connection to the legacy databases.

6.4.5.2.7 Install and fully test the target environment (Part 6)

An open target environment with appropriate installation tools should be identified, selected, installed and fully tested.

6.4.5.2.8 Create and install the necessary gateways (Part 7)

Gateways should be designed, created and installed to ensure consistency and accuracy in replicating the legacy system's functionalities in the target system and in transferring electronic document-based information. Gateways typically have two roles. One is to insulate selected components from changes being made to others. The other is to act as a translator of requests and data among mediated components. Gateways should be carefully tested with samples of legacy document-based information to ensure consistency and accuracy.

6.4.5.2.9 Migrate the legacy database (Part 8)

The legacy database should be migrated to the target database.

6.4.5.2.10 Migrate the legacy applications (Part 9)

The legacy applications should be migrated to the target applications.

6.4.5.2.11 Migrate the legacy interfaces (Part 10)

The legacy interfaces should be migrated to the target interfaces. Legacy interfaces (e.g. character-based menus and screens) are likely to be replaced with graphical interfaces.

7 Developing a long-term preservation strategy

7.1 Long-term preservation policy

Storage repositories seeking to provide long-term preservation of authentic and processable electronic document-based information should create material describing its policies and procedures. This documentation will enhance the authenticity of electronic document-based information as evidence and help to ensure it is handled consistently and uniformly. In addition, it will help a storage repository prepare for questions of trustworthiness that typically are raised in legal proceedings.

A best practice long-term preservation strategy's policy should contain the following elements:

— a section stating that providing long-term preservation of authentic and processable document-based information is a goal of the storage repository and the identification of other repository goals and responsibilities;

— a description of the type of custody that the storage repository undertakes for electronic document-based information, e.g. legal or physical;

— a description of the electronic document-based information management best practices to which the storage repository adheres;

— identification of the circumstances under which migration activities will be undertaken and the methods and rationale for such activities;

— an explanation of the types of compliance audit that will take place;

— clarification of the roles of storage repository personnel and a description of any responsibilities that are outsourced.

7.2 Quality control

Electronic document-based information preserved in accordance with established rules and procedures is generally considered to have greater authenticity and ultimately better credibility in legal proceedings.