Functional Requirements for a Preservation System

1 Introduction

1.1 Background and scope

The Digital Preservation Testbed project has investigated potential long term preservation strategies for different types of digital records, namely e-mail messages, text documents, spreadsheets and databases. For each of these record types a recommendation has been produced on the most appropriate preservation strategy. These recommendations address the question of preservation mostly at the level of individual records.

The present document discusses the functional requirements for a preservation system to implement the recommendations. Functional requirements relate to the things that the system should do, as opposed to non-functional requirements, such as the necessary capacity of the system, development methods or performance characteristics.

In this document the main features and functions of a preservation system and the choices that must be considered when designing such a system are discussed. Preparing a list of individual, precise, traceable requirements that could be used for building or procuring such a system is beyond the scope of our work.

It is assumed that we are dealing with a system for storage and preservation of digital archival records. Much of the document is also relevant to other kinds of preservation system, for example systems for preserving scientific or historical data.

All the activities required at each stage of the life of a record, from its creation, through its maintenance by the creating department, transfer to a long-term archive and long term maintenance of the record in the custody of the long-term archive are discussed. All stages of this process are essential to the long term preservation of records. Example activities associated with a record are illustrated below in figure 1.

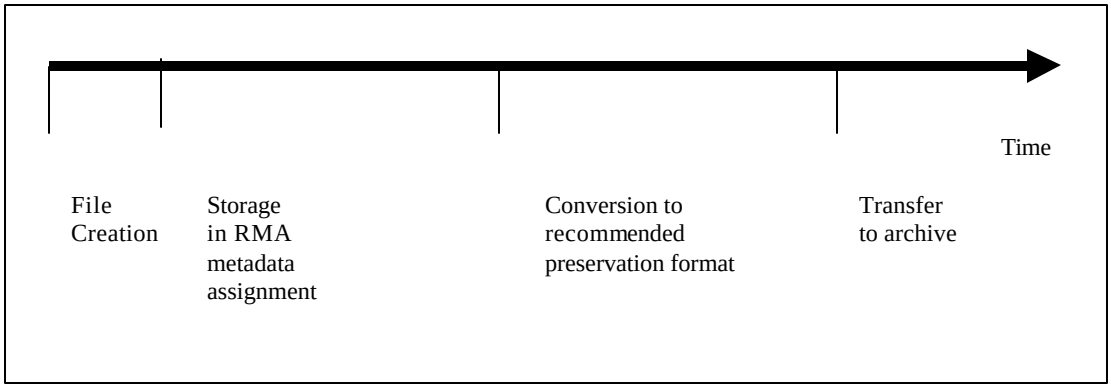

Figure 1 – The activity timeline of an example record.

1.2 Context

This document integrates individual requirements associated with a range of long-term preservation strategies into a single set of functional requirements. An essential part of a preservation system is the repository where

records are held and managed. The DEPOT 2000 project produced a "Functional Design for a digital depot"[1], setting out in detail the functions and data structure for such a repository system. The DEPOT 2000 document deliberately sets the choice and implementation of a long-term preservation strategy to one side. In Chapter 5, we discuss how the present preservation system requirements relate to the requirements for a repository. This document does not explicitly consider the costs associated with a digital archive. That is addressed in a separate Testbed publication.[2]

These functional requirements are closely related to the recommendations of the Testbed project on the best approach for the long term preservation of records of various types. Recommendations have been produced for the preservation of e-mail messages, text documents, spreadsheets and databases. The recommendations themselves concentrate on the actions required to preserve an individual record. This document considers the computing environment and systems required to enable these recommendations to be implemented effectively and efficiently.

The Reference Model for an Open Archival Information System (OAIS Model) is generally regarded as the standard work for defining a framework and procedures for the preservation of digital records. This document focuses mainly on the "Preservation Planning" element of the OAIS model, but we also consider other aspects, notably how to deal with the requirement to provide and maintain sufficient representation information. The relationship between the Testbed recommendations and the OAIS model is discussed further in Chapter 5.

Whilst still in the custody of the organisation that created them, records are likely to be held in a Records Management Application (RMA)[3]. In Chapter 3, we refer to published requirements for such systems and comment on how the need for long-term preservation affects these requirements.

1.3 Definitions

The Testbed glossary has a comprehensive list of definitions used in this document. However, a few of the terms are of particular importance and so are defined here, to set the scene for the rest of the discussion.

Digital Record – a digital entity, preserved in the form of a file assembly, that an archive receives and preserves. The archive preserves the file assembly of the record using a strategy dictated by the record type. For simplicity this archived file assembly is referred to as the record in this document, though the file assembly is intrinsically dependent upon suitable applications to read and represent it authentically.

Long-term digital archive – a system designed and used to receive and store authentically preserved digital records. The meaning of "long term" depends on context, but in this document this usually refers to a period of greater than 10 years.

2 Records continuum

The concept of a records continuum is very useful for our discussion. It can be defined as:

"...a consistent and coherent regime of management processes from the time of the creation of records (and before creation, in the design of record keeping systems), through to the preservation and use of records as archives."[4]

---

1 *DEPOT 2000. Functional Design for a Digital Depot. Nico van Egmond, Hans Hofman, Jacqueline Slats, Tamara van Zwol.*

2 *See the Testbed Cost Model report.*

3 *An alternative widely used expression is Electronic Records Management System (ERMS)*

4 *Australian Standard AS 4390-1996:* Records Management, *Part 1:* General, *Clause 4.22*

In our discussion of the preservation of digital records, there are three important phases associated with the continuum, see figure 2:

Before accession to the long-term archive
The accession process
Long-term preservation in the long-term archive

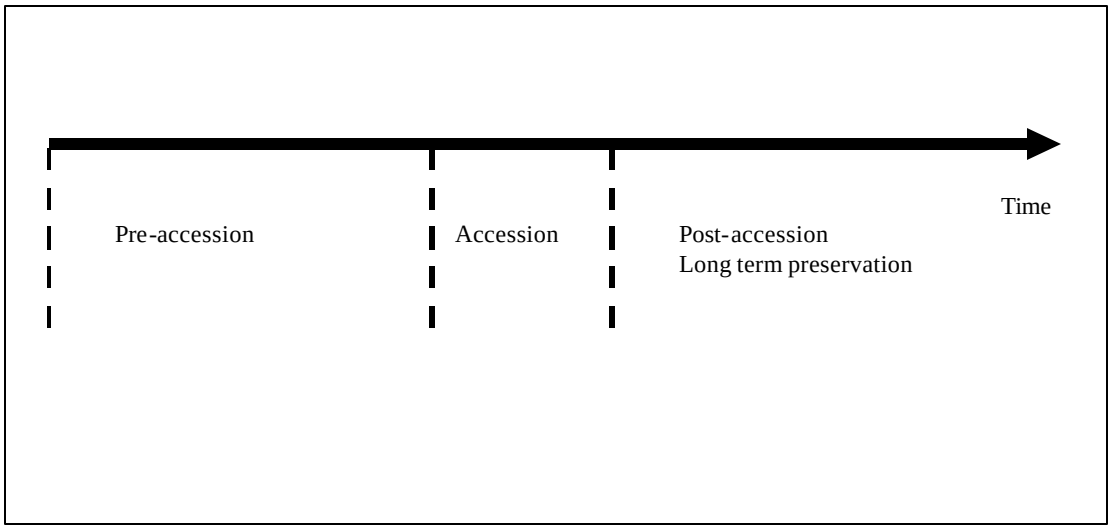

Figure 2 – The 3 phases of a record continuum.

These three phases will be described in the following three chapters. Each phase of the records continuum is equally important: if a break in the "consistent and coherent regime" occurs before the records are in the care of the long-term archive, then their preservation is put at risk.

Many of the requirements applying to the third phase are also relevant for the first phase.
This is because in many cases the digital records will be in the custody of the creating organisation for a period long enough for the need and application of preservation actions to prevent digital obsolescence.

3 Before accession to archive (phase 1)

It is a requirement that records should be carefully managed according to well-defined procedures. In most cases, a RMA will be a useful tool in achieving this aim. One likely exception to this is in the case of databases. Although the RMA may be a good place to store documentation and metadata about databases that have been identified for preservation, it will usually make sense to store and maintain the database itself outside of the RMA.

3.1 Selecting an appropriate RMA

A number of organisations have published requirements for RMA systems. These include:
European Commission: Model Requirements for the Management of Electronic Records (MoReq)[5]
Softwarespecificaties voor Records Management Applicaties voor de Nederlandse overheid (Remano)[6].
UK Public Record Office[7]
US Department of Defense[8]

These requirements documents give details of the features which RMA software should have. The UK Public Record Office and the US Department of Defense each have a programme to evaluate software packages against their requirements. The outcome of this work is lists of compliant software systems[9][10]. A less formal but nonetheless useful list of products and their features is maintained by the Dutch government Elektronische Overheid project[11].

3.2 Configuring an RMA

Most off-the-shelf RMA software products are designed to be flexible and highly configurable. Indeed, because every organisation is different, such systems need significant configuration if they are to be able to implement the records policies of each organisation.

The features of an RMA which must be configured to suit an organisation include:
the records classification scheme based on business processes (see ISO 15489-1)
the set of metadata items which are associated with each record
retention schedules and disposal actions

In addition, the Testbed recommendations for the preservation of records state further specific requirements for the organisation in relation to the data and associated metadata: in some cases there is a need to transform files from one format to another. Some of these could be implemented within the framework of an RMA, whereas for others, it makes more sense to use external tools. These activities are discussed in more detail in the next section.

3.3 Capturing records and preparing them for preservation

3.3.1 Use of a Records Management Application

An important requirement for the reliable preservation of digital records is that, as soon as possible after their creation and their identification as a record which should be retained, they should be placed into a well-organised and controlled storage environment, i.e. a records management application. MoReq refers to this as the "capture" of the record in the RMA.
At this point the record should be classified: that is, there must be a defined classification system for the records held by the organisation. On capture of the record in the RMA, the place in this classification where the record belongs should be identified. In some cases, the capture and classification can take place automatically as part of a workflow. If the main workflows of the organisation can be integrated with the RMA, then much of the work of records capture can be automated, reducing the burden on the users and reducing the risk of errors or omissions.

---

5 *http://www.cornwell.co.uk/moreq*

6 *Softwarespecificaties voor Records Management Applicaties voor de Nederlandse overheid (Remano). Hans Waalwijk, Geert-Jan van Bussel, Peter Horsman, Archiefschool Versie 4.12 9 september 2002 http://www.digitaleduurzaamheid.nl/bibliotheek/docs/remano_versie4_12bis.doc. This is based on MoReq, but has been adapted for the specific requirements of the Dutch government.*

7 *http://www.pro.gov.uk/recordsmanagement/erecords/2002reqs/default.htm*

8 *http://jitc.fhu.disa.mil/recmgt/standards.htm*

9 *UK PRO list of approved systems: http://www.pro.gov.uk/recordsmanagement/erecords/2002reqs/2002listofapprovedsystems.htm, http://www.pro.gov.uk/recordsmanagement/erecords/1999reqs/1999listofapprovedsystems.htm*

10 *US Department of Defense list of approved systems: http://jitc.fhu.disa.mil/recmgt/*

11 *http://www.digitalegereedschapskist.nl/systemen*

3.3.2 Capturing metadata

On capture of the record in the RMA, the necessary metadata should be added to the record. It is beyond the scope of this document to specify which items of metadata are required. This may depend on the particular requirements of each organization. However, the previously referenced MoReq and US Department of Defence documents, for example, provide good starting points. As far as possible the RMA should be configured to collect metadata automatically. Items such as the name of the user submitting the record and the date and time can easily be captured by software. If the record is created as part of a programmed workflow, then metadata identifying the business process and the place of this record within that process can also be captured by the RMA. Note that different types of record may have different metadata requirements.

Clearly, an essential requirement for any records management system is that the association between a record and its metadata must be maintained with 100% reliability.

In addition to the metadata required for registering the institutional context, there may also be additional items required for technical or preservation reasons, for example relating to the format of the computer file(s) making up the record or describing any transformations that have been carried out on the record. The Testbed recommendations for each record type discuss preservation metadata requirements in more detail.

3.3.3 Conversion of file formats

It is the responsibility of the creating organisation to conduct file format conversions of records to formats specified in archival regulation 'Geordende en toegankelijke staat archiefbescheiden'.

Some file formats pose a higher risk for long-term preservation than others, so for many records it may be necessary to transform the original computer files into new formats, which have been chosen for their suitability for long-term preservation. There are several examples of this in the Testbed recommendations. In general, it is best to carry out this transformation as soon as possible after the record has been created and identified for long-term preservation. This will cause additional storage requirements, but we contend that these are outweighed by the increased reliability and trustworthiness of carrying out the transformations whilst the original processing software is still operational and knowledge of the original purpose and context of the record is still available in the creating organisation.
Organisations are not obliged to apply conversion to file formats recommended for long term preservation for records with a short retention schedule. Organisations do have the obligation to keep these records on good and ordered state. Note that, even for records with a short retention schedule, there may sometimes be a need to perform preservation actions such as migration to maintain access to the record.

Conversion to the recommended preservation format could be incorporated as a function of an RMA. When a record is first entered in the RMA and associated with a retention schedule involving long-term preservation and transfer to the National Archives, then automatic conversion tools could be invoked by the RMA to convert the file to the required new format. This process should also involve evaluation of the success of such transformations - these should be automated where possible, but in some cases may also involve an element of manual checking. If a record is initially assigned a short retention schedule, but later it is decided that the record should be transferred to the National Archives (and so preserved indefinitely) then the transformation to the chosen preservation format should be associated with the act of modifying the retention schedule. More information about the design and testing of preservation approaches and automated testing procedures is given in section 5.10.

3.4 Transfer from one RMA to another

The current archiving regulations in the Netherlands state that records must be transferred to the National Archives within 20 years of their creation. It is unlikely that the organisation responsible for the records will use the same RMA throughout this period. Like other areas of information technology, new products and approaches appear frequently and at some point there may be benefits for an organization to change from one RMA to another.

When this occurs, the records contained in the RMA must be transferred reliably to the new system. In some cases, the system manufacturers may include specific facilities for transferring between specific combinations of systems, but it is not generally safe to rely on such features being available.
To ensure that this process of migration between RMAs can be successfully achieved, it should be a requirement that the RMA can export and import records in a vendor neutral interchange format, which maintains the logical structure linking the components of a record and also maintains the links between records, for example grouping

records into dossiers, or links between a record containing a particular file format and another object containing representation information for that format.

When selecting or building an RMA in the present day, it is not possible to know if the export format of the RMA will match the supported import formats of future RMAs. However, if the system uses a well-documented non-proprietary export format, then it should be possible without excessive effort to provide a matching import facility in the future. XML is likely to be a suitable basis for such an interchange format. MoReq includes a useful section on "Transfer, export and destruction" of records, requiring that an RMA 'must provide a well-managed process to transfer records to another system or to a third-party organisation'.

From a technical point of view, the process of transferring records from an RMA to the National Archives has much in common with the transfer between one RMA and another. However, from a regulatory and record-keeping point of view there are additional aspects to consider in the former case and these are discussed in the next chapter.

4 Accession (phase 2)

4.1 Overview

The accession stage of the records continuum covers the transfer of the records from the creating organisation to the long-term archive. Dutch archiving regulations [12] place a number of conditions on how the records being transferred to the Nationaal Archief should be organised, including for example the file formats and types of metadata required.

The transfer itself could be carried out over a communications network or on removable media. We can define a group of related records transferred to the long-term archive at one time as an *accession*. The accession will typically involve a large number of records, each with associated metadata. The records will normally be organised into dossiers or other groups and may have other links between them. These are the same issues as discussed in Section 3.4. The transfer format must be able to represent the basic information and all the important links between items.

4.2 Processing the transferred records

Each record must be assigned a unique reference on being added to the archive and this becomes the definitive reference. The records may already have a unique reference assigned during the pre-accession stage. This should also be retained.

To minimise the human effort required at this stage, one or more records transfer formats should be developed and published by the long-term archive, working in consultation with the record-creating organisations. The transfer format is a specification to allow the grouping of files and metadata together as a package. This can then be implemented as an export format in the RMAs of the record creating organisations, allowing transfer of records to the National Archives to take place as simply as possible.

As the records are imported to the long-term archive, any related cataloguing systems should be updated as required, based on the record metadata.

At this stage it may be possible to extract further automatic metadata about the files and records being submitted. This may not be necessary if the creating organisation has already ensured that all required metadata are already present. The accessioning system must verify that all necessary metadata have been provided, according to whatever metadata schema has been agreed between the National Archives and the creating organisation. The contents of each computer file must also be checked for viruses and the result of this check stored in the record metadata.

---

[12] *"Geordende en toegankelijke staat van archiefbescheiden" http://www.nationaalarchief.nl/images/3_2598.doc*

5 Preservation in the archive (phase 3)

5.1 Introduction and background

This chapter discusses the functions and features a digital archive should have to support the preservation process.

The Consultative Committee for Space Data Systems (CCSDS) has published a recommendation for a "Reference Model for an Open Archival Information System (OAIS)".
The OAIS model defines an information model for a long term preservation system and lists a set of responsibilities that the system should fulfil. It has now been adopted as ISO standard 14721:2002. It is therefore a useful basis for our discussion and a few important points from the OAIS model are summarised briefly here. For a full explanation of OAIS, the reader is referred to [OAIS ref][13].

5.1.1 OAIS Model

Users

The OAIS Model defines three main types of external users of a digital archive: producers, consumers and management. The producers are those creating the information to be preserved, the consumers are those making use of the preserved information and the management is responsible for high-level policy making for the archive (day-to-day administration of the system is defined as one of the functions of the OAIS).

Data Structure

The OAIS defines an Information Object as a Data Object combined with Representation Information. The Data Object can be thought of as the computer file or files making up the object. Because computer files or bitstreams in isolation cannot be meaningfully interpreted, there must be additional representation information, in the form of documentation or computer software.

There may be several layers of representation information and some items of representation information may be shared by more than one Data Object. For example, one element of representation information is likely to be a specification of the mapping from bits to characters in a particular data object (for example following the Unicode standard). Since a large number of records will have this information in common, it does not make sense to store it separately for every record.

The OAIS Model points out that the amount and type of representation information required depends on the intended users of the Information Object, known in the OAIS Model as the "Designated Community". A Designated Community has an associated Knowledge Base, that is a set of knowledge which a member of that community can be assumed to have. The Representation Information must be sufficient for a member of the Designated Community to be able to understand the Information Object.

Functions

Figure 3 illustrates the main groups of functions of an archival information system and we reproduce it here for convenience.

[13] *OAIS. http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html*

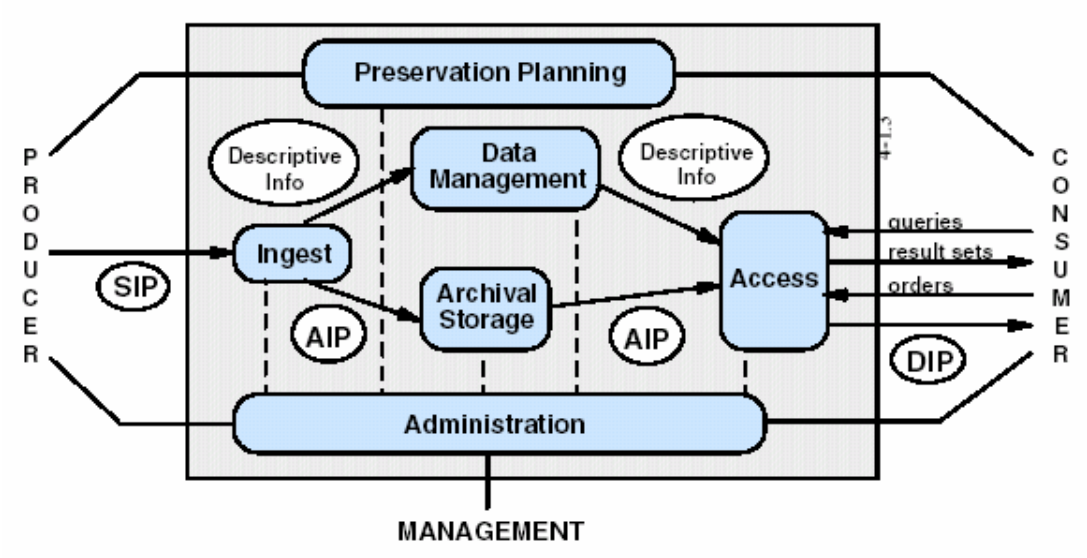

Figure 3 - OAIS Functional Entities, reproduced from the CCSDS Recommendation

From this it can be seen that the main groups of functions that a preservation system must implement are:
Ingest
Data Management
Archival Storage
Access
Administration
Preservation Planning

Of these functional groups, most of the functions are associated with the repository system, and as explained above, it is only intended to discuss in this document those aspects with a particular relevance to long-term preservation. In the following sections, reference will be made to the related OAIS function to show how our discussion fits into the OAIS framework.

The abbreviations SIP, AIP and DIP appearing in the diagram above stand for Submission Information Package, Archival Information Package and Dissemination Information Package and refer to the way the information is organised during ingest to the archive (SIP), while stored in the archive (AIP) and when being distributed to users (DIP). Refer to the OAIS Model for more detailed information.

Thus the OAIS Model provides a framework for long-term preservation issues and discusses some of the approaches. In the rest of this chapter, we will elaborate on the requirements for a system to provide long-term preservation.

5.1.2 Existing approaches

There are few examples of operational digital archiving systems and even fewer with well-developed preservation strategies. However, a number of organisations have carried out pilot studies, made initial implementations of preservation approaches, or have published information on their preferred future strategies. A selection of these are briefly reviewed in this section.

UK National Archives

The UK National Archives have an operational Digital Archive (since April 2003), storing UK government records and making them available to the public[14]. The focus of this system at present is on a secure repository for digital records (metadata and content files). They have not yet fixed on a particular long-term preservation strategy but have ensured that the system is designed to allow future strategies to be effectively implemented.

The policy of the UK National Archives is to accept files in any format. In addition to the core digital archive itself, they have a database of file format information, known as PRONOM[15]. PRONOM is a system for "managing information about the file formats used to store electronic records, and the software applications needed to render these formats" and forms a component of their technology watch programme.

The Digital Archives at the UK National Archives includes the concept of multiple "manifestations" of a record, that is different digital representations of the same record, providing a framework to enable possible future migrations.

Public Record Office Victoria (PROV)

The Victorian Electronic Records Strategy (VERS) programme of PROV[16] has developed a strategy for long-term preservation based around a choice of a limited number of preservation formats. In contrast to the UK National Archives, they insist that electronic records are submitted to them in one of a small defined set of allowable file types. For 'printable' record types, they require that the record creating organisation submits files to PROV in PDF format. By limiting the file formats in the repository to a small carefully chosen list, PROV aim to simplify future preservation effort.

Another key element in the VERS approach is to store records as self-contained digital objects. They have defined an XML format which combines their metadata schema with the files themselves. The file (or files) containing the record content are base64 encoded (to convert them from binary format to text format) and are stored as elements in the XML document. This has the benefit of minimizing the amount of supporting IT infrastructure needed to allow the record to be correctly reconstructed and interpreted, as no external system I needed to maintain links between record metadata and content files. However, such an external system is still required to support efficient and controlled access to the records.

PROV plan to have an operational digital archive around the end of 2005.

NARA

At the time of writing, NARA are involved in the procurement process for their Electronic Records Archives (ERA) project [17]. An important element of their proposed approach for ensuring long-term accessibility of digital records is the conversion of files to *persistent formats*. The ERA requirements document explains this as follows:

"A persistent format is one that is supported by a preservation strategy for diminishing the impacts of technological obsolescence, minimising dependence on specific hardware and software and enabling retrieval and output of authentic copies in the future. An ideal persistent format would be self-describing and be able to be validated in accordance with open, non-proprietary standards".

[14] *http://www.pro.gov.uk/about/preservation/digital/archive/default.htm*
[15] *http://www.pro.gov.uk/about/preservation/digital/pronom/default.htm*
[16] *http://www.prov.vic.gov.au/vers/welcome.htm*
[17] *http://www.archives.gov/electronic_records_archives/acquisition/draft_rfp.html*

5.1.3 Repository functions versus preservation functions

As discussed above, this document concentrates on the long-term preservation functions of a digital archive. In addition there are many essential functions of a digital archive which we are assuming will be provided. These are discussed in detail in other documents, for example the DEPOT 2000 publication of the Rijksarchiefdienst and they are summarised very briefly below.

The basic function of the repository is to manage records and their metadata, including relationships between records. It must do this extremely reliably: the system must include comprehensive back-up and recovery systems so that records cannot be lost, even in the case of a disaster.
There must also be systematic monitoring and replacement of storage media, to ensure that data on disks and tapes is not corrupted. Access to the system must be controlled, to regulate what activities can be carried out by which users or types of users.

The repository must deal with accession of records and access to records. These aspects are influenced by long-term preservation issues and so are discussed in more detail elsewhere in this document: see Chapter 4 and Section 5.11.

5.2 Manifestations

For a migration based strategy, it is necessary for the preservation system to include the concept of multiple representations or manifestations of a record. As Thibodeau explains [18] it is possible to represent digital records in different ways without losing authenticity: as long as the conceptual record, as presented to the user, is sufficiently similar then it is possible and acceptable to change the file format and rendering software. He gives the example of a text document that can be represented as either an MS Word file or a PDF file.

The OAIS Model refers to various types of migration. Migration including a format conversion is one of these migration types and is referred to as Transformation in the OAIS context. Transformation of a record (or AIP) leads to a new AIP which is a new Version of the previous AIP. The previous version of the AIP will usually be retained, at least temporarily, so that the transformation process can be checked, or as a source format for possible further transformations.

Each manifestation should include or link to all information needed for the record (e.g. shared representation information would typically be linked rather than copied), so that it should not be necessary to access more than one manifestation in order to render the record authentically. A manifestation may involve multiple computer files.

Users of the system would not necessarily have access to all of the retained manifestations. There could be one preferred manifestation for user access, or possibly more than one alternative. This mechanism could also be used for redaction, where original and redacted copies of a record would be different manifestations.

Each time a new manifestation is added, information about the process used to create it must be recorded in an audit trail.

5.3 Technology Watch and Preservation Planning

The OAIS Model identifies Preservation Planning as a group of functions that a digital archive system must support. This group is broken down into:
Monitor Technology
Monitor Designated Community
Develop Preservation Strategies and Standards
Develop Packaging Designs and Migration Plans

The Monitor Technology function is often referred to as a Technology Watch. This is one of the most important functions within preservation planning and the one that we discuss in most detail here. This involves monitoring

[18] *Ken Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years" http://www.clir.org/pubs/reports/pub107/pub107.pdf.*

the various technological components used in the digital archive, to identify as early as possible if a component may be in danger of becoming obsolete. Early identification of potential problems allows action to be taken before access to archive contents is made difficult or impossible.

One approach to this is to monitor the file formats held in the archive and to maintain information on the status of each file format, for example which application software is able to understand and render the file format, on which hardware platforms the application software is available, whether the application software is still supported by its manufacturer and so on. The UK National Archives have produced a software system called PRONOM intended to help tackle this problem[19].

In the DNEP[20] system of the Koninklijke Bibliotheek, this concept has been formalised as the Preservation Layer Model[21]. Each item in the deposit system will be associated with one or more View Paths, defining the application software, operating system and hardware required to access the item. Note that one record in an archiving system may consist of multiple computer files in multiple formats. Files in different formats will in general have different view paths, so there may be more than one view path per record. This implies that the technology watch must take place at the level of the computer file, not just at the level of the archival records.

When it is identified that a record in the system is in danger of becoming inaccessible, then some kind of preservation action must take place, for example migrating it to a new format, or creating an emulation of a hardware environment where the original software application can continue to run. The steps that will be involved in this are discussed in Section 5.10.

5.4 Requirements for e-mail preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving e-mail messages[22]. The recommendations advocate converting an e-mail message to an XML document, linked to separate files containing the message body, in plain text or HTML format or both, and any message attachments. Because e-mail attachments can be any type of file, the e-mail recommendations do not make specific recommendations on how to deal with attachments, beyond noting that an appropriate preservation strategy for that file type should be applied. Using the view path concept discussed in Section 5.3, each computer file in the record will need its own view path. For example, the XML may be viewed using simple text viewing software.

Processing the e-mail message to transform it to the required preservation format (i.e. this set of linked files plus metadata) should be carried out as soon as possible after the message has been identified as requiring preservation. This could be incorporated as a modification to the e-mail application being used, or it could be implemented as a feature of a records management application.

The Testbed recommendation means that the e-mail record is represented by several files, linked in a particular way. The preservation system must have the ability to make and maintain these links between the files within a record. In the e-mail recommendations it is suggested that this can be done by having an XML file at the core of a record, defining the relationships between the component files and hence the structure of the record. This is one, quite flexible, approach to achieving this objective. Other solutions are also possible, for example using a relational database.

If one or more of the files within a record must be updated as part of its preservation strategy, then this should require creation of a new manifestation of the whole record. This is a general feature for all the record types described in the next few sections.

---

19 *http://www.pro.gov.uk/about/preservation/digital/pronom.htm*

20 *http://www.kb.nl/kb/resources/frameset_kb.html?/kb/ict/dea/index-en.html*

21 *"The Long-Term Preservation Study of the DNEP project – an overview of the results", Raymond J. van Diessen, Johan F. Steenbakkers, IBM Netherlands, Amsterdam, ISBN 90-6259-154-X*

22 *http://www.digitaleduurzaamheid.nl/bibliotheek/docs/bewaren_van_email.pdf (Dutch) or http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-email-en.pdf (English)*

5.5 Requirements for text document preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving text documents[23]. The Testbed recommendation is to keep the original file, together with a PDF and optionally also an XML representation.

Due to the use of multiple file formats there is a need for multiple representations of same record. The XML approach consists of several associated files including a XML content file, a XSLT (XML Stylesheet language transformations) stylesheet and related image files. The XSLT should transform the XML into a XSL-FO (XSL formatting objects) file. A suitable XSL-FO processing application is required to render the resulting XSL-FO file. The PDF approach consists of a single binary file that can be viewed using PDF viewing software.

Processing the text document to transform it to the required preservation form at (i.e. this set of linked files plus metadata) should be carried out as soon as possible after the text document has been identified as requiring preservation. This could be incorporated at document creation, post creation using a separate application, or it could be implemented as a feature of a Records Management Application.

As with e-mail messages, there must be a mechanism for representing the relationships between the different files making up the record, which could be the appropriate use of XML or a database.

A technology watch is required to identify major changes in either the PDF format or relevant XML implementations. Such changes will require a re-evaluation of the present strategy.

5.6 Requirements for spreadsheet preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving spreadsheets[24]. The Testbed recommendations advocate keeping the original file as well as converting the spreadsheet to an XML document assembly.

The XML approach consists of several associated files including a XML content file, optionally an XSLT stylesheet and related image files. There is not a requirement to actively render the XML files, rather they can be viewed in text form using a simple text editor.

Processing the spreadsheet to transform it to the required preservation format (i.e. this set of linked files plus metadata) should be carried out as soon as possible after the spreadsheet has been identified as requiring preservation. This could be incorporated at creation, post creation using a separate application, or it could be implemented as a feature of a Records Management Application.

There must be a mechanism for representing the relationships between the different files making up the record, which could be the appropriate use of XML such as a linking file, or a database.

5.7 Requirements for database preservation

This section examines the specific features of a preservation system required to support the Testbed recommendations for preserving databases[25]. The Testbed recommendation is to keep the original database data file(s) or export file(s), an XML representation of the application database tables and associated information in the underlying database and documentation to describe key input and outputs associated with the application. This documentation will contain SQL and code associated with the transactions described.

The XML approach consists of creating several associated files including a XML overview file, and XML content files representing each of the application tables. The application documentation can be preserved and viewed according to the text document recommendations. There is not a requirement to actively render the XML files, rather they can be viewed in text form using a simple text editor. However, there is the possibility to conduct a second conversion to transform the XML contents files of the database tables into an active database.

[23] *http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-textdocs-en.pdf*
[24] *http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-spreadsh-en.pdf*
[25] *http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf*

Processing the database to transform it to the required preservation format (i.e. this set of linked files plus metadata) should usually be performed when the database ceases to be active or when the database must be transferred for long-term preservation. This could be done using a separate application, or it could be implemented as a feature of a Records Management Application.

There must be a mechanism for representing the relationships between the different files making up the preservation object. There are several possibilities to implement this, for instance the 'framework approach' described in chapter 4 of the Database recommendations [26].

5.8 Other record types

The Testbed project investigated e-mail messages, text documents, spreadsheets and databases. Testbed is therefore not in a position to make specific recommendations about other types of digital records. However, similar principles will apply. The types of application software available to render or interpret a file format should be considered. File formats whose specification is clearly defined and publicly available are to be preferred over closed proprietary formats. If the chosen strategy is of conversion to a format more suitable for long-term preservation, then careful testing and evaluation of the conversion process must be undertaken to ensure that the essential characteristics of the original records are retained.

5.9 Preservation actions

The most common type of 'preservation action' is likely to be conversion of the file or files making up the content of a digital record from one format to another. When this occurs, it is recommended that the original files, as submitted to the archive, are retained. If multiple file format conversions take place over a period of time, a decision must be made whether to keep intermediate versions. Keeping all versions gives the maximum protection against loss of information, but at the cost of increased storage and management requirements.

When such a file format conversion, or other preservation action, is carried out, it is essential that sufficient information about the conversion process is collected so that the provenance and authenticity of the new representation of the file can be established. It must be possible to determine which file was the predecessor of the newly created representation and to understand what process was applied to make the conversion. If digital signature or hashing approaches are used to check for corruption of files or metadata, then these must be re-applied to the new format, once the necessary quality checking has been carried out.

When the format of the computer files making up a record is changed, there is a risk that information is lost or changed, potentially leading to a loss of authenticity. Therefore, when such a migration takes place, it is important to put in place a thorough checking procedure. Checking can be implemented by use of a separate automated testing 'module' which can be frequency enhanced in the future to systematically reduce dependency on manual testing.

A separate test system should be developed to conduct on-going research into preservation approaches, for example, migrations to a new format, or an emulation of a hardware environment where the original software application can continue to run. The research should keep track of, and examine relevant new technology. The research scope should also cover the development of automated testing techniques. Promising approaches need to be thoroughly tested to ensure that the authenticity of a record is maintained and it is properly checked. The results of this research are updates for the preservation and automated testing modules in the preservation system and guidelines specifying the most suitable approach to preserve a range of record types.

When a technology watch, described in section 5.4, identifies that a group of records in the system is in danger of becoming inaccessible then some kind of preservation action must take place. The whole group of records must be classified first by record type and then by sub categories grouped by matching authenticity requirements.

For each record group the most suitable preservation approach should be identified, which ensures the authenticity requirements are met. For migration the selected approach should ideally be an open standard,

---

[26] *http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf*

widely used and well documented. The decision will be guided using the results generated from research conducted using the test system described above, and the Testbed long-term preservation recommendations.

For each record group the identified preservation approaches should be applied. Apart from the use of emulation, such preservation action will lead to new manifestations of each record (see section 5.3). The preservation action should be recorded in the metadata and the preservation log of each record.

5.10 Representation information

There are often cases when many records require the same representation information, such as an XML schema, or a reference to a standards document, such as the Unicode standard. In such cases the information does not need to be explicitly stored with each record, but rather an appropriate linking approach should be used so that only a single instance of the representation information needs to be stored. A number of possible linking methods may be adopted depending upon the nature of the digital archive system implemented, but consideration should always be given to the ease with which links can be maintained if the records are transferred in future to a new and potentially different digital archive system.

5.11 Security

Security is a very important aspect in a trustworthy digital repository. One of the difficulties of digital record keeping is the ease with which digital records can be changed or deleted and a digital archive system must be designed to minimise risks of this type. The standard approach to IT system security is to assign users to particular roles. Different roles are assigned different privileges, defining what they are allowed to see, whether they are allowed to add new entries, or edit or delete existing entries.

The UK National Archives approach has been to implement two systems – a closed system, which is the definitive version of the archive and an open system, available to the public, containing copies of the records in the closed system. This means that even if the security access controls to the open system are compromised, there is no risk of unauthorised access to the closed system. The closed system is separated by an 'air gap' from the open system and is not linked to any external network.

The security model should allow for some records to be defined as available for public viewing and others not available. The organisation must decide whether it will allow users to see if closed records exist, even if they are not permitted to view the contents. There will typically be a requirement for redaction, where users can see edited versions of sensitive records with confidential information removed.

To allow verification that record contents or metadata have not been accidentally or deliberately changed, digital signatures or hash functions can be used. The VERS approach advocates applying a digital signature to both the record content and metadata. If changes or additions to the metadata are made, then the modified version must be resigned.

The system must maintain an audit trail, recording all additions, changes or deletions, noting which user was responsible and when it occurred.

5.12 Import/export requirements

A digital archive system should be designed to allow the export of records and associated information to users or archivists requiring this information. Note the export is a copy of the original record required. The digital archive system should further be designed to allow large scale transfers of records from one archive to another. This can be achieved by the use of suitable functionality built into the archive system combined with the use of a transfer file format (preferably an open standard and easy to access, such as XML). The discussion in section 3.4 of the database recommendations[27] is also relevant here and describes the export requirements in more detail.

[27] *http://www.digitaleduurzaamheid.nl/bibliotheek/docs/volatility-permanence-databases-en.pdf*

5.13 Metadata structure

Metadata may be associated with a group of records, individual record, manifestations of a digital document or computer files. Groups of related records in an archive are commonly organised using a dossier or similar organisational system. Such structures should be defined in the metadata. It is important that metadata functionality is implemented to allow flexibility in an archival system. The approach should allow for the possibility of metadata templates changing, due for example, to new required elements being added.

5.14 Alternative preservation approaches

Testbed has recommended preservation approaches for four record types: email, text documents, spreadsheets and databases. These recommendations are variants of migration or conversion to standards. The choice of approach has been influenced by what is feasible and verifiable at the present time. It is recognised that alternative preservation approaches may be developed and implemented in the future. Because of this, the preservation system should be designed so that introducing a new preservation approach has the minimum effect on the other parts of the system (as described in section 5.10).

Emulation (more specifically the 'software-emulation-of-hardware approach') as a preservation approach has been considered by the Testbed project, though not recommended at this stage, as it remains largely unproven in a digital preservation context. If emulation is implemented as a preservation approach in the future there would be differences to the way the preservation and access modules of the digital archive system operate. There would not be a need to alter the original files associated with a record, but more complex information about the software and hardware environment to view the record would be required, as well as operational copies of the necessary application and operating system software.