

IBM
long

KB
term

preservation

study

010010

preservation
requirements
in a deposit
system

10





KB

010010 preservation 10
requirements
in a deposit
system

Dr. Raymond J. van Diessen

01001011

01000010

Design: Steven L. Stijger
Published by: IBM Netherlands, Amsterdam
IBM / KB Long-Term Preservation Study
Report Series Editor: Dr. Raymond J. van Diessen

Available from:
IBM External Communications
PO Box 9999
1000 CE Amsterdam
The Netherlands

Koninklijke Bibliotheek
PO Box 90407
2509 LK The Hague
The Netherlands

Title: **Preservation Requirements in a Deposit System**

ISBN: 90-6259-156-6
Author: Dr. Raymond J. van Diessen
Date: December 2002
Copyright: IBM / Koninklijke Bibliotheek

*This study was commissioned by the Koninklijke Bibliotheek,
National Library of the Netherlands*

01 IBM / KB 010000010

long-term preservation study

The National Library of the Netherlands (Koninklijke Bibliotheek, KB) is faced with the problem of preserving large amounts of digital documents for the long term. These documents come from two sources: from media published directly in digital form and from digitizing paper documents. In 2000, the KB and IBM started building an electronic deposit system ("Digital Information Archiving System or DIAS"), the technical core of the infrastructure for KB's e-Deposit for the Netherlands.

From the beginning it was clear that this project could not rely on out-of-the-box solutions alone because up to that time no solution readily addressed both the aspects of large volume and durable storage as well as the long-term preservation requirements. So an IBM / KB Long-Term Preservation Study (LTP Study) was initiated as part of the overall project of developing an electronic deposit system.

The primary objective of the LTP Study was to investigate the functionality required for the long-term preservation (hundreds of years) of the digital information stored in DIAS. This study has resulted in 6 reports: one overview report and five specific reports, each one addressing an important aspect of long-term preservation in its own right.

Participants in the LTP Study:

IBM

Raymond J. van Diessen
Raymond Lorie
Sidney Huiskamp
Hans Verhoeven

Koninklijke Bibliotheek

Johan F. Steenbakkens
Titia van der Werf-Davelaar
Patricia Alkhoven
Adriaan Lemmen

RAND Corporation

Jeff Rothenberg

British Library

Deborah Woodyard

I would like to thank all the participants for their input and enthusiasm. The results make an important contribution to the development and implementation of dedicated functionality for the long-term preservation of digital information and for guaranteeing long-term access.

Report Series Editor,
Raymond J. van Diessen

Titles of the Reports Series

Number 1: The Long-Term Preservation Study of the DNEP Project - an Overview of the Results

This report explains the reasons and objectives behind defining the LTP Study as part of the overall project to implement an electronic deposit system. It also provides a quick and general overview of all the study results, which are then elaborated on in the other published reports.

Number 2: Authenticity in a Digital Environment

Authenticity acquires a new meaning in a digital context. Normally objects are physical and their physical characteristics are the main source for defining authenticity. Moreover, authenticity is not a single concept, but involves different aspects that can be associated with an object:

- ∄ A traceable path from the object's origin to its current ownership.
- ∄ Measures and techniques for safeguarding against and/or recognizing modifications.
- ∄ Techniques for establishing the use of original materials.

The problem of digital objects is that in fact they are just conceptual objects. A digital object is a conceptual object to be interpreted (rendered) by executing the digital object in a specific IT infrastructure (hardware & software). This report focuses on defining a framework in which we can define what is actually meant when one speaks of an authentic digital object.

Number 3: Preservation Requirements in a Deposit System

The initial DIAS release only provides basic functionality for preserving and rendering the stored digital objects for the long term. One of the primary responsibilities of the LTP Study is to define the functional requirements of the Preservation Subsystem, which is scheduled for development later. This report identifies requirements of the DIAS Preservation Subsystem so as to provide the services and functions for monitoring the technical environment associated with the digital objects stored in DIAS.

The Preservation Subsystem can be summarized by the following three objectives:

- ∄ Identifying digital objects that are in danger of becoming inaccessible because of changes in technology.
- ∄ Implementing the activities associated with technical preservation.
- ∄ Supplying the requisite technical metadata in order to generate / validate the environments needed during digital object delivery.

Number 4: The UVC: a Method for Preserving Digital Documents - Proof of Concept

Within IBM Research in Almaden, Raymond Lorie was already working on a combined emulation / migration approach to preserve a certain class of digital objects with an approach called the Universal Virtual Computer (UVC).

The main idea consists of archiving a program P along with the data file that decodes the data and returns the information to a future client based on a logical view. The logical view of the data is simple and self-contained enough to be interpreted without any specific software or hardware. Program P is written for the Universal Virtual Computer (UVC) that is general, yet basic enough to continue to be relevant in the future. Given the simplicity of the UVC, it will be relatively easy to write an emulator of the UVC in the future on a real machine of that time. The emulated machine will run the program P and return all data in an easy to understand logical view of the data.

The LTP Study conducted a proof of concept with the KB to test the UVC approach in a library environment. The PDF format was selected because it is the primary data format for electronic publications to be stored in DIAS.

Number 5: Managing Media Migration in a Deposit System

Storage technology obsolescence makes media migration a necessity. Data has to be copied from one storage medium to another on a regular basis. However, the fact that storage technology becomes obsolete is not the only trigger for rewriting previously stored digital objects. All storage media degrade over time and have to be rewritten either on the same medium (refreshing) or on another medium (migration).

Ordinarily media refreshment / migration would be a straightforward process. However, the large amounts of storage associated with an electronic deposit system introduce certain volume-specific requirements. Most electronic deposit systems define their storage capacity needs in several TeraBytes (10^{12} Bytes). Take a deposit system with 100 TeraBytes of information stored on tape, for example. Let's assume that you want to migrate all this information to an optical storage medium. Current optical storage media have a capacity of around 5 GigaBytes and a write speed of around 4 MegaBytes/second. A quick calculation shows that a complete migration to optical storage would take at least 290 days (100 TeraBytes / 4 MegaBytes per second)!

This report describes the actions to be taken to manage media migration / refreshment effectively within an electronic deposit system, focussing specifically on the media migration issues within DIAS. Potential additional capacity required for media migration might be created by redundancy and parallelism.

Number 6: Archiving Web Publications

More and more Web publications are becoming a primary source of information and will thus be stored as digital objects in DIAS. Web publications have specific characteristics and requirements that DIAS must meet if they are to be archived successfully.

This report investigates the issues and requirements introduced by archiving Web publications and their potential impact on DIAS.

01 contents



1/	Summary	1
2/	Objectives	3
	2.1 DIAS within OAIS Context	3
3/	Preservation Layer Model	7
	3.1 Technical Metadata	7
	3.2 Preservation Layer Model	11
	3.3 View Paths	14
	3.4 Initial DIAS Version	16
	3.5 Installed Electronic Publications	18
	3.6 Preservation of Technical Metadata	19
4/	Preservation Process Model	21
	4.1 Publication Ingest	21
	4.2 Creating new PLMs and View Paths	24
	4.3 Monitor Technology	25
5/	Summary	27
	Appendix A: References	31
	Appendix B: Glossary	33
	Appendix C: DIAS Release 1 Recognized File Types	35
	Appendix D: Possible PLM Implementation	37
	Appendix E: Use Cases	39



01 summary



This report presents part of the results of a study conducted by IBM for the National Library of the Netherlands (Koninklijke Bibliotheek, KB). IBM was commissioned to conduct the Long-Term Preservation Study in parallel with the development and implementation of a system for the 'Depot van Nederlandse Elektronische Publicaties (DNEP). IBM delivered the first version of the electronic deposit system, called Digital Information Archiving System (DIAS), this autumn. The design of the deposit system is based on the OAIS functional model [CCSDS 2001]. The functionality required to preserve the digital objects for the long term was not included in detail in the requirements for the initial version of DIAS. Hence at the time of the Request for Proposal and the possible awarding of the contract to IBM, the precise requirements for long-term preservation could not be defined appropriately.

As a deposit library the KB is facing the problem of preserving large quantities of digital documents for the long term. These documents originate from two sources: digitized paper documents and new media types published in digital form. The digital form offers many advantages but it also involves a challenge of its own. The issues are how can we insure that such documents (digital files) will be preserved for a long time - surviving changes in storage technology, computer hardware, software, formats, etc. - and how can we guarantee the accessibility in the future.

The term digital object is used to identify all digital information stored in the broader multimedia context, including images, sound, video and even programs. Preservation of digital objects needs to be examined from at least three perspectives: medium preservation, technology preservation and intellectual preservation.

Medium Preservation

The medium on which the information is stored can decay. Medium preservation focuses on preserving the medium on which information is stored, such as tapes, magnetical disks, optical disks, CD-ROMs and the like. Backup is appropriate, as is copying to other devices of the same type, a technique known as "refreshing".

Technology Preservation

The rapid changes in the storage formats and in the software that renders the electronic information present an even bigger problem than medium decay. We need to be aware of technology obsolescence as an even more challenging problem than medium decay, and undertake appropriate steps to address this problem (technology preservation). Rather than simply refreshing, we also need to speak of migration and emulation:

- € **Migrating** information to new technology / format stages as they become available and as the old technologies / formats cease to be supported by vendors and the user community.
- € **Emulating** old and obsolete technologies / formats on current technology platforms.

Intellectual Preservation

There is yet a third preservation requirement, intellectual preservation, which addresses

the integrity and authenticity of the information as originally recorded. Preservation of the media and the associated rendering software will only address part of the need for preserving the digital object. The need for intellectual preservation arises because of the ease with which an identical copy can be made quickly and flawlessly. This is perhaps the greatest asset of a digital object, but is also its greatest liability. Changes, which are largely undetectable, can be made to the content just as easily as making a copy.

This report will focus primarily on technology preservation aspects, although medium and intellectual preservation are important aspects in their own right. Within the OAIS context this is referred to as Preservation Planning:

"Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base."

The DIAS implementation and the Long-Term Preservation Study will refer to OAIS' Preservation Planning module as the DIAS Preservation Subsystem.

This report describes the requirements of the Preservation Subsystem as an extension of the basic preservation functionality provided by the initial DIAS version. Developing the deposit system in parallel with defining the Preservation Subsystem requires that the individual components of the system be loosely coupled through pre-defined interfaces. This proven design principle was applied throughout the entire DNEP implementation project.

Investigations into the requirements of the Preservation Subsystem made it apparent that no individual structure describing technical metadata would suffice. Taking the long-term challenges seriously, no individual structure could be sufficiently general or could be expected to be applied through a series of major technology changes. Fortunately some general principles could be identified which collectively provide the foundation on which to support specific technical metadata and the functionality required by the Preservation Subsystem.

2/ 01 objectives



The Preservation Subsystem provides the services and functions for monitoring the technical environment associated with the digital objects stored in DIAS. It provides recommendations for ensuring that the digital objects stored in DIAS remain accessible over time, even if the original computing environment becomes obsolete. And it supports the availability of functionality for rendering the information in the digital objects.

The Preservation Subsystem can be summarized by the following three objectives:

- € Identifying the digital objects in danger of becoming inaccessible due to technology changes.
- € Implementing the activities associated with technical preservation, i.e. implementing migration and emulation strategies.
- € Registering the technical metadata needed to generate / validate the software and hardware environments required to render the digital object.

The Preservation Subsystem can be summarized by the following three objectives:

- € Identifying the digital objects in danger of becoming inaccessible due to technology changes.*
- € Implementing the activities associated with technical preservation, i.e. implementing migration and emulation strategies.*
- € Registering the technical metadata needed to generate / validate the software and hardware environments required to render the digital object.*

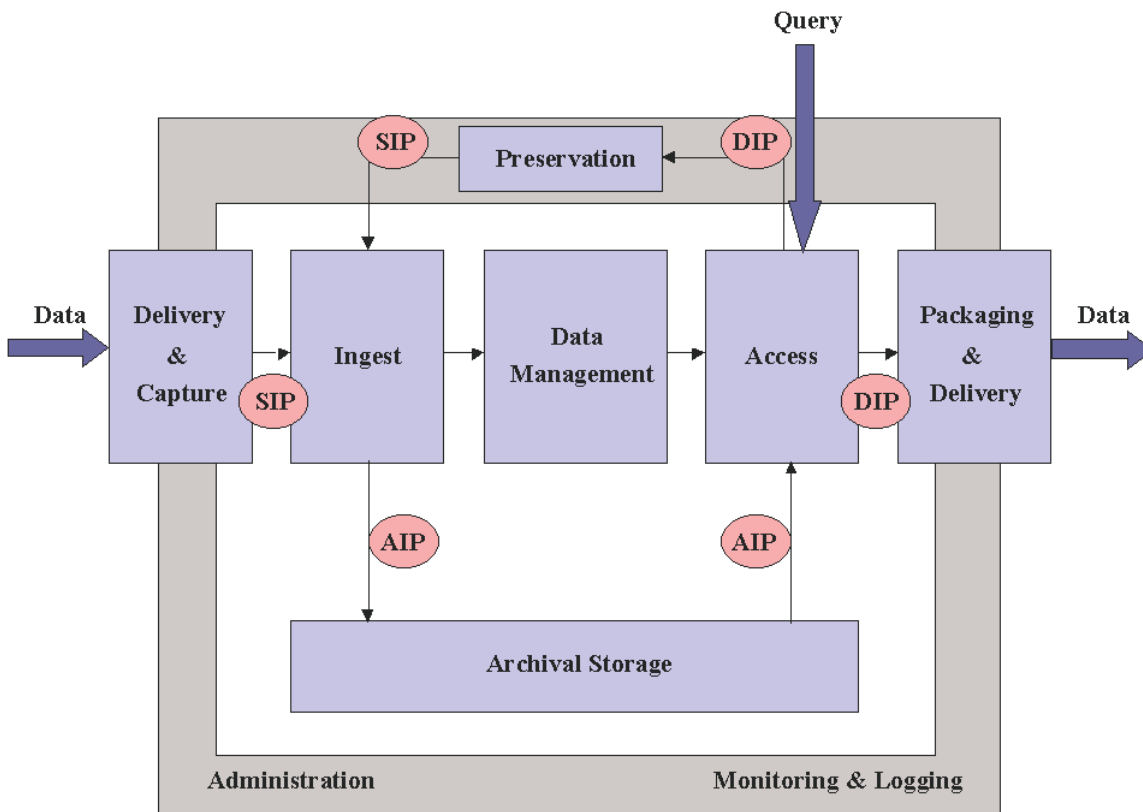
2.1 DIAS within OAIS Context

The current DIAS implementation conforms to the Open Archival Information System (OAIS) standard [CCSDS 2001]. The revision of July 2001 introduces a new module, Preservation. See Figure 2.1. Until this version of OAIS appeared, no additional

functionality had been identified to support technical preservation efforts.

The top-level functional components of an electronic deposit system are:

- € **Ingest**
 Receives a digital object prepared by an interfacing process (Delivery&Capture) and loads it into archival storage.
- € **Archival Storage**
 Takes care of storing and retrieving the digital objects and ensures the integrity of the bit-streams.
- € **Data Management**
 Takes care of storing and retrieving metadata associated with the digital object.
- € **Access**
 Makes the archived digital object and its associated metadata available through an interfacing process (Packaging&Delivery).
- € **Administration**
 Is responsible for the operation of the system.
- € **Monitoring & Logging**
 The whole workflow for handling the electronic publications is monitored and its quality is controlled.
- € **Preservation**
 Is responsible for the long-term access and readability of electronic publications.



2.1 / OAIS overview

The two electronic deposit interface components (Delivery&Capture and Packaging&Delivery) provide the external interfaces for the electronic deposit:

- € **Delivery&Capture** takes care of the pre-processing of digital objects to be ingested. It receives or captures digital objects and offers a workspace for verification based on the specifications for ingestion into the electronic deposit.
- € **Packaging&Delivery** is the output interface for the deposit system. It handles the post-processing of digital objects retrieved from the electronic depository. It negotiates access requests, delivers and installs electronic publications, along with appropriate viewing or operating software and metadata, for direct access by the requestor.

The primary difference between the DIAS design and the OAIS reference model is the positioning of the technical metadata associated with each digital object. OAIS uses the concept of an Information Package as a container for two types of information called Content Information and Preservation Description Information (PDI). See Figure 2.2. The Content Information contains the actual content or what we call the digital object. The PDI contains all the technical metadata required to render the digital object. Together the two components of the Information Package provide the information needed to render the digital object.

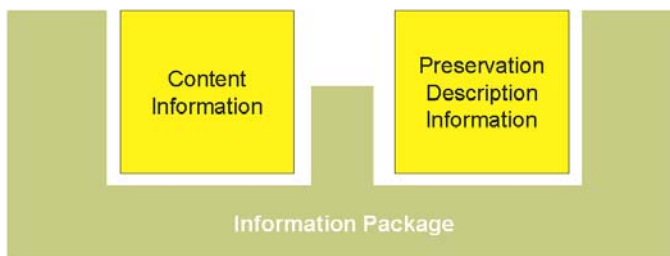


Figure 2.2 / OAIS Information Package

DIAS physically stores the technical metadata associated with specific digital object types in the Preservation Subsystem. One objective of the DIAS Preservation Subsystem is to register the technical metadata needed to generate / validate the software and hardware environments that are needed to render the digital object.

The separation of the Content Information and the Preservation Description Information is based on functional coherence and flexibility arguments. Separating the digital object from the technical metadata enables DIAS to update and manage the technical metadata without modifying the archived digital object. This is being achieved by linking the separate managed objects through the use of a FileTypeID. Any (virtual) rendering tools are stored and handled as distinct digital objects. The following Chapter will elaborate more on the exact implementation of this approach.

In other aspects the DIAS Preservation Subsystem does not differ substantially from the functionality of the preservation module as defined within the OAIS functional reference model. The Preservation Subsystem needs to be aware of technology obsolescence that could render a digital object inaccessible and take appropriate steps to prevent this (technology preservation).

The Preservation Subsystem is also responsible for the actions to be taken when technology changes threaten to make a digital object inaccessible. Two basic categories of activities can be distinguished:

- € **Migrating** information to new technology / format stages as they become available and as the old technologies / formats cease to be supported by vendors and the user community.
- € **Emulating** old and obsolete technologies / formats on current technology platforms.

It is likely that the deposit system will use both techniques to guarantee that digital objects stored in its deposit can be rendered at any time in the future in the technological environment being utilized at that time.

The separation of the Content Information and the Preservation Description Information is based on functional coherence and flexibility arguments. Separating the digital object from the technical metadata enables DIAS to update and manage the technical metadata without modifying the archived digital object.

3/

01 Preservation Layer Model 000010

The objective of the Preservation Subsystem is to register information about the IT environment (hardware and software) that is needed to render the digital object:

- € Identifying the digital objects that are in danger of becoming inaccessible due to technology changes.
- € Implementing the activities associated with technical preservation, i.e. implementation of migration and emulation strategies.
- € Registering the technical metadata needed to generate / validate the software and hardware environments required to render the digital object.

Hence the heart of any Preservation Subsystem is the technical metadata associated with the digital objects. Digital objects will be rendered useless over time if no information regarding the IT infrastructure requirements is managed. The IT infrastructure knowledge can either be lost or technology changes can make the particular IT infrastructure obsolete.

This chapter introduces the concept of a Preservation Layer Model (PLM) to describe the structure of the technical metadata needed by the Preservation Subsystem to implement its objectives. A layered description of the preservation metadata is also used in OAIS [Lupovici and Masanès 2000].

3.1 Technical Metadata

Digital objects are rendered by a combination of hardware and software. The digital object itself is merely an abstraction represented by a string of digits containing zeroes and ones. The software, i.e. the application software, creates the context by which the digital object can be interpreted. The actual hardware running the software provides the physical representation; without both the digital objects as such would be meaningless. Figure 3.1 shows the components involved in the rendering process.

Within the rendering process shown beneath four general layers can be identified, which are involved in the rendering of any digital object:

1. **Data format layer:** the format defines the structure of the bit stream, i.e. the intangible digital object.

2. **Application layer:** software applications are used to create, modify and view information in its intended format.
3. **Operating system layer:** the operating system provides the shared functionality, such as interfacing to peripherals and file management, which is needed by every application.
4. **Hardware layer:** the hardware is the computer platform on which the intangible digital object is rendered into real physical objects, such as a printed document or a screen representation.

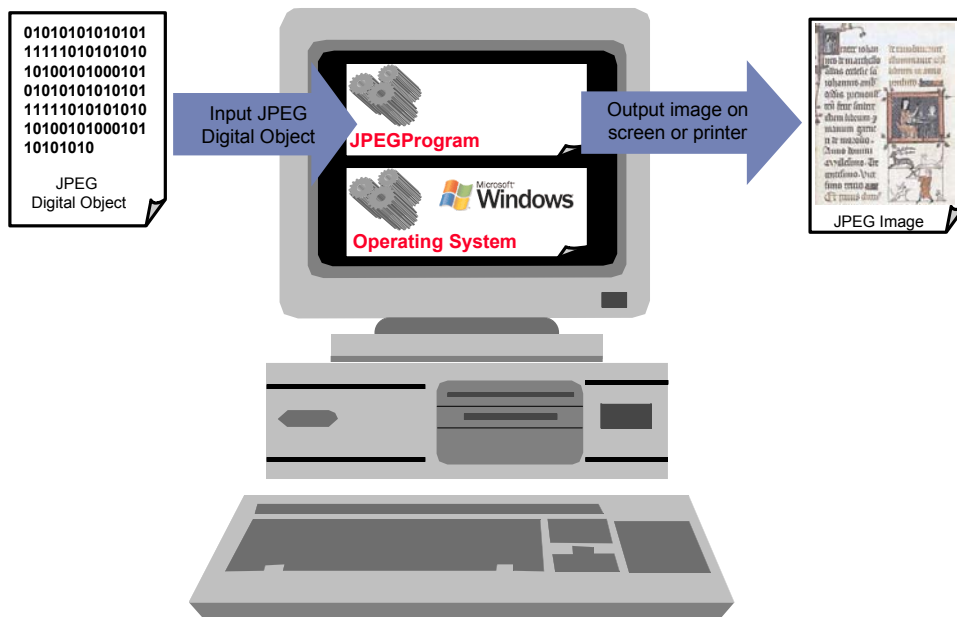


Figure 3.1 / Rendering a digital object

This layered model shows the chain of software and hardware dependencies digital objects carry with them. It is called the Preservation Layer Model (PLM) because it is used as a tool to manage the preservation of the rendering process. Over time the chain is broken in different places, as software and /or hardware components become obsolete. One broken link is sufficient to make a digital object unusable.

The layered model applies to all digital objects, irrespective of their specific format and usage. This suggests that it can be used generically. The generic structure can be used for several purposes:

- ∄ As a basic tool to map technical dependencies between computer hardware, operating systems, application software and file formats.
- ∄ To assess the longevity of a given file format or application software.
- ∄ To identify digital objects that are in danger of becoming inaccessible because of obsolescence in one of the layers.
- ∄ To generate possible View Paths for rendering a digital object.

In particular, it can be used as a meaningful framework for assessing preservation strategies. In addition, a deposit's authenticity criteria will influence the technical metadata maintained for specific digital objects.

As part of the IBM / KB Long-Term Preservation Study, the report "Authenticity in a Digital Environment" introduces an authenticity framework for defining the authenticity aspects of a digital object in relation to its interpretation process [Diessen and Werf-Davelaar 2002]. For each of the interpretation schemata defined in the authenticity framework the organization maintaining the electronic deposit has to decide whether or not it views that particular interpretation schema as part of the information that has to be preserved. These decisions will, in turn, influence the type of technical metadata that needs to be maintained by the Preservation Subsystem.

Therefore the exact technical metadata maintained by the Preservation Subsystem will depend on the specific context: digital object type, application support, operating system, hardware platform and the deposit's authenticity objectives and capabilities.

3.1.1 Interdependencies

A quick scan of some real case examples showed that no single data structure would suffice to define the technical metadata for a variety of digital objects. Different combinations of digital object types, applications, operating systems and hardware result in different structures for the rendering environment.

The example in Tables 3.1 and 3.2 shows how the technical metadata differs for the Lotus Domino server application. In this case the hardware requirements will vary depending on the operating system on which the application is run. This situation is not

Platform	Windows 95/98	Macintosh	Windows NT
Certified operating system versions	Windows 95; Windows 98	Mac OS 9	Windows NT Workstation 4.0
Processors supported	Intel Pentium	PowerPC	Intel Pentium
RAM	8 MB minimum 32 MB or more recommended	32 MB physical, 64 MB virtual minimum; 64 MB physical, 80 MB virtual recommended	16 MB minimum 32 MB or more recommended
Disk space The minimum amounts are the disk space required for installing default files. More disk space is required if databases are replicated locally or copied locally.	Notes client: 69 MB minimum 112 MB or more recommended Designer client: 70 MB minimum 236 MB or more recommended Administrator client 78 MB minimum 182 MB or more recommended	75 MB minimum; 100 MB or more recommended (standard client) 75 MB minimum; 150 MB or more recommended (designer client)	Notes client: 69 MB minimum 112 MB or more recommended Designer client: 70 MB minimum 236 MB or more recommended Administrator client 78 MB minimum 182 MB or more recommended
Monitors supported	Color monitor required	Color monitor required, 256 colors or greater.	Color monitor required

Table 3.1 / Hardware requirements for a Lotus Domino Server

uncommon and is also one of the arguments for separating technical metadata from the actual archived digital object. Recall the difference compared to the OAIS reference model described in Chapter 2 that positioned the technical metadata alongside the digital object in one Information Package. All the differences in values of technical metadata resulting from different hardware and operating system combinations have to be maintained inside the Information Package. The information also has to be attached to every information package.

Linux patch requirements	AIX patch requirements	Windows service packs
<p>Identifying required patches for Linux is difficult because Linux is distributed as different levels, with different packages for an application in each level. However, by setting requirements based on kernel and library levels, then evaluating each distribution based on this, a set of requirements can be determined.</p> <p>A certain set of patches can be guaranteed by certifying and supporting distributions. At the lowest level, Linux kernel 2.2.5 or greater is required, along with glibc 2.1.1 or greater, and libstdc++ 2.9.0 or greater. Each of the supported/certified distributions contains these levels or higher. There is one exception to this rule: the version of glibc/libstdc++ installed must contain the libstdc++-libc6 naming convention. If it does not, you must make the appropriate link yourself (for example, on Red Hat 6.0, the file is /usr/lib/libstdc++-libc6.1-1.so.2. on Caldera 2.2, this naming convention is not used and thus you must link /usr/lib/libstdc++-libc6.1-1.so.2 to the file /usr/lib/libstdc++_so.2.9.0, which is the appropriate library). A distribution that meets these requirements should be able to accommodate the Domino server.</p>	<p>AIX 4.3.1 was certified with Domino R5 using the patches listed below. Although individual operating system patches and service packs are not certified, Lotus realizes that there are later patches that may become publicly available after testing, and acknowledges that these more recent updates may fix additional problems. You may wish to apply these newer patches as they become available. You can obtain and install the following patches from IBM or download them from IBM's Web site: http://www.ibm.com</p> <p>Fix IX85874 bos.adt.debug 4.3.1.1 bos.adt.syscalls 4.3.1.1 bos.rte.bind_cmds 4.3.1.1 bos.rte.commands 4.3.1.1 bos.rte.control 4.3.1.1 bos.rte.cron 4.3.1.1 bos.rte.install 4.3.1.1 bos.rte.shell 4.3.1.1 bos.sysmgmt.smit 4.3.1.1 devices.graphics.com 4.3.1.1 X11.base.rte 4.3.1.1 X11.compat.lib. X11R5 4.3.1.1 xIC.rte 3.6.4.1 Fix IY01777IY06473 (Needed to prevent the server from crashing when running NSD).</p>	<p>R5 was initially certified on Windows 95/98 and Windows NT 4.0 with the following Service Packs applied. Although individual operating system patches and Service Packs are not certified, Lotus realizes that there are later service packs that may become publicly available after testing, and acknowledges that these more recent updates may fix additional problems.</p> <p>You may wish to apply these newer service packs as they become available. You can obtain and install the following patches from Microsoft Corporation or download them from Microsoft's Web site: http://www.microsoft.com/msdownload/</p> <p>Windows 95/98: Service Pack 1 and Service Pack 1 Updates Windows NT 4.0 Workstation and Server: Service Pack 4</p>

Table 3.2 / Patch instructions for Lotus Domino Server

The flexibility presented above makes it impossible to define one single format to structure the specification of the technical metadata for individual digital object types. We can only establish general areas, which are an important source for technical metadata and acknowledge the dependencies among the different layers, see Figure 3.2.

Current operational specification might not always be viewed as a minimum requirement. Extrapolation to higher "versions" may not always result in a workable environment. For example, increases in hardware performance may result in timing problems for applications designed and implemented for a "lower" hardware platform. What about the effects of new versions designed to improve the performance rather than to drastically change the functionality; would the digital object still be rendered correctly? To be accurate not only a lower limit have to be established but an upper limit as well in order to address the fact that improvements can invalidate the proper operation of the application and/or operating system.

Another issue is the support of runtime extensions, for example Open DataBase Connectivity (ODBC) drivers. Different applications might require different versions of these extensions, which might not always be installed at the same time. Unrelated extensions could also influence each other and produce errors in otherwise properly functioning components when installed separately.

Further research into the effective specification of technical metadata needed by the Preservation Subsystem will be required.

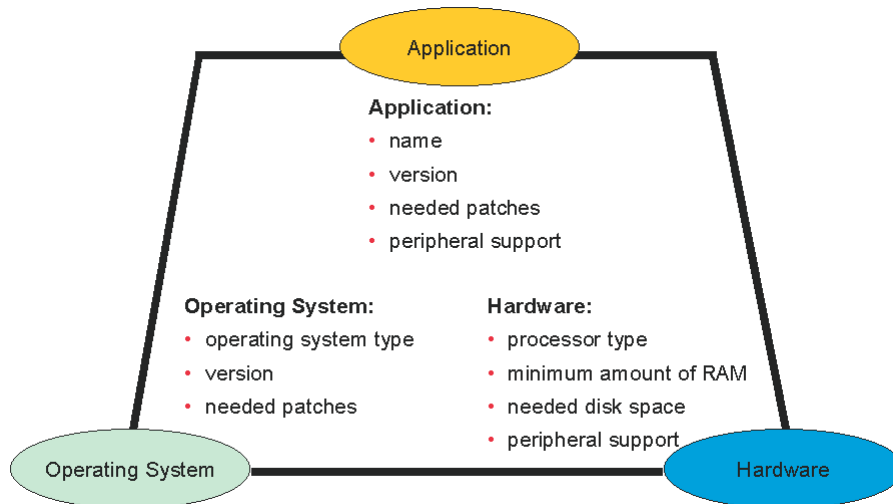


Figure 3.2 / Technical metadata definition dependencies among the different layers

3.2 Preservation Layer Model

The experience from initial test data has shown that it would be hard to define one single globally applicable View Path template that supports all technical metadata requirements even in the current technological environment.

A new concept, the Preservation Layer Model (PLM), is introduced to represent the different template structures used to steer the specification of the technical metadata needed and maintained by the Preservation Subsystem. Section 3.3 shows that a specific PLM depends on the digital object type, the operating system, the hardware, and the peripheral devices. Technology changes over time will also introduce demands for additional PLMs.

Thus the LTP Study concluded that the objective should not be to define a single PLM or rather a 'View Path' template, but should rather focus on the implementation of a flexible mechanism for defining PLMs on demand and for efficiently managing the associated View Paths. A View Path represents a full set of functionality needed to render the information from a digital object.

The experience from initial test data has shown that it would be hard to define one single globally applicable View Path template that supports all technical metadata requirements even in the current technological environment.

The most common computer architecture in use today is the Von Neumann architecture. A computer based on this architecture can be viewed as a series of layers, each of which builds on the functionality provided by the previous layer. Figure 3.3 shows that most modern computers involve roughly six general layers [Tanenbaum 1990].

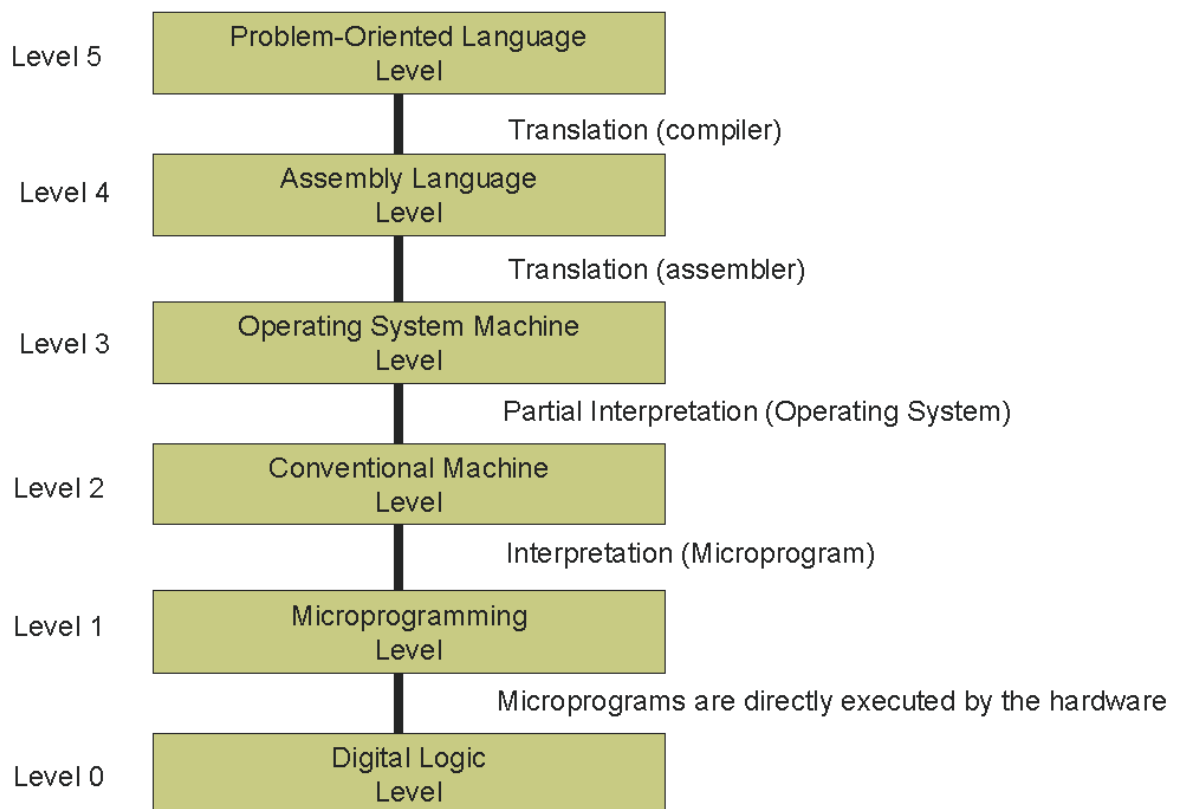


Figure 3.3 / Contemporary computer architecture

Within our PLM metadata model we build on this concept of layering to describe the rendering IT infrastructure. The PLM metadata model provides the constructs by which

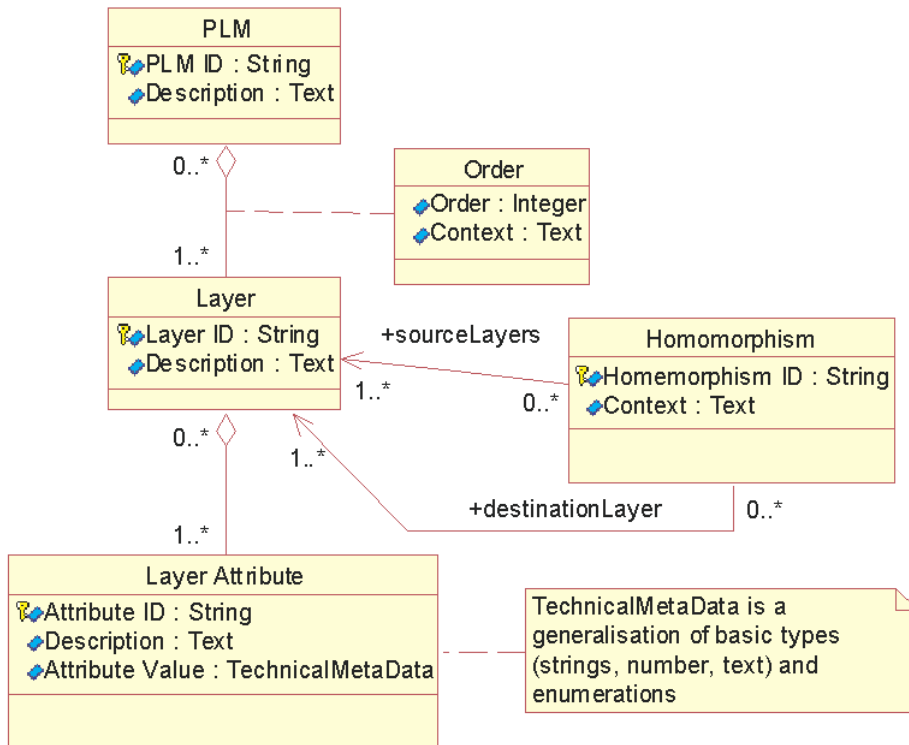


Figure 3.4 / Preservation Layer Model (PLM) metadata model

individual PLM models can be defined. Furthermore, the PLM metadata provides the ability to relate different layers of different PLMs to one another using homomorphic relationships. Figure 3.4 shows the Unified Modeling Language (UML) class model [Booch et al. 1999] for the PLM metadata model.

PLMs are defined by a set of ordered layers. Each layer in turn can use a set of defined technical metadata attributes to describe the characteristics that specific layer focuses on. An attribute can either be a simple type such as an integer, a string, text or an enumerated user defined type like an Operating System (Windows 3, Window 95, Windows 98, Windows ME etc). Both layers and attributes can be reused in different PLMs. The Order class describes the order and context of a layer in a specific PLM.

One of the most frequently stored digital object types within the KB's electronic deposit is actually PDF. Current PDF implementations do not require any specific hardware support in order to display the content. The basic technical metadata can be specified by a three layer PLM, see Figure 3.5. The right side of the figure shows a specific View Path for PDF files and the left side shows the PLM metadata. Note that the layer attribute Version in the layer Application has been defined as a collection of versions in order to provide the means for displaying the range of versions that can be used to render a specific PDF document type.

To avoid overloading an identification component (attributes, layers, PLMs) with other semantically meaningful information, the semantics of a component are stored in a separate description attribute associated with each component. The sole purpose of the IDs is to uniquely identify a component, nothing more and nothing less.

Application Layer (3)
 Application name: String
 Version: Collection of String

Operating System Layer (2)
 Operating System: (Windows, MacOS)

Hardware Layer (1)
 Processor Architecture: (Intel Pentium, PowerPC)

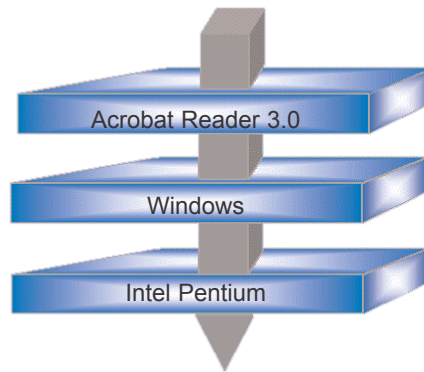


Figure 3.5 / PLM example

Within the first implementation we will use 25-character string attributes to identify the components. Approximately 150 characters can be used at each position resulting in an address space of roughly $2.5 \cdot 10^{54}$ possible component IDs for each type. Even if 1000 new components are introduced every second, it still will take approximately $2.5 \cdot 8 \cdot 10^{43}$ years for a component type to run out of address space. We believe that this is an acceptable solution looking at long term preservation.

The most common computer architecture in use today is the Von Neumann architecture. A computer based on this architecture can be viewed as a series of layers, each of which builds on the functionality provided by the previous layer.

3.3 View Paths

As mentioned before, a View Path is a full set of functionality for rendering the information contained in a digital object. With the aid of the PLMs the Preservation Subsystem maintains different View Paths for all the digital object types stored in DIAS. Every digital object type has at least one View Path specifying an IT infrastructure capable of rendering the digital object.

The link between the AIPs stored in DIAS and the View Paths to be stored in the Preservation Subsystem is established through the Table of Contents (ToC). Every AIP has a XML formatted ToC that specifies the name and file type of every file (digital object)

stored in that AIP. The attribute FileTypeId is a number that references a certain file type - version combination. Appendix C shows the known file types, which are recognized in the initial version of DIAS.

The attribute FileTypeId is the only reference mechanism the Preservation Subsystem will use to associate technical metadata with digital objects stored in DIAS as an AIP. This provides a high degree of flexibility and actually eliminates the need for DIAS to update AIPs once they are ingested into the system. The fact that the AIP cannot be modified after it has been ingested into the system provides a level of quality control. Figure 3.6 show the UML class diagram for the connectivity between View Paths and AIPs by FileTypeId.

A suggested detailed design based on a relational database management system is discussed in Appendix D.

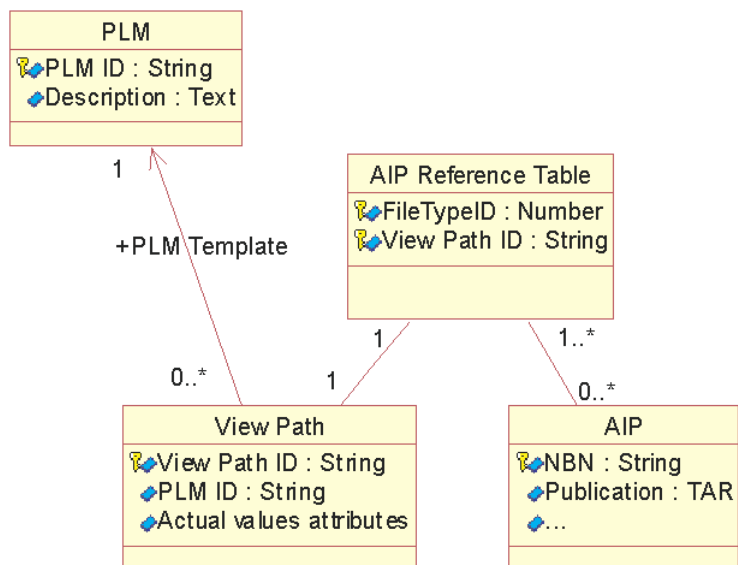


Figure 3.6 / AIP reference table

The attribute FileTypeId is the only reference mechanism the Preservation Subsystem will use to associate technical metadata with digital objects stored in DIAS as an AIP. This provides a high degree of flexibility and actually eliminates the need for DIAS to update AIPs once they are ingested into the system.

3.4 Initial DIAS Version

The initial release to be delivered by IBM will not yet contain the Preservation Subsystem. One of the main reasons for initiating the LTP Study in parallel was to gain sufficient knowledge regarding the issues involved surrounding long-term preservation to be able to specify a first element for the Preservation Subsystem.

This introduced a problem for the developers of the initial DIAS version. Although it was agreed that there would be no Preservation Subsystem, some preservation functionality is required to make DIAS workable. A preliminary solution had to be developed that performed the preservation requirements in the absence of the Preservation Subsystem. During this intermediate DIAS phase all the digital objects stored in the deposit system can be accessed on a specific KB defined Reference Platform.

The Reference Platform infrastructure is a set of computers whose configuration is completely controlled by the KB. The Reference Platform is a complete combination of installed software and hardware needed to render the digital objects. The components, represented in Figure 3.7, are:

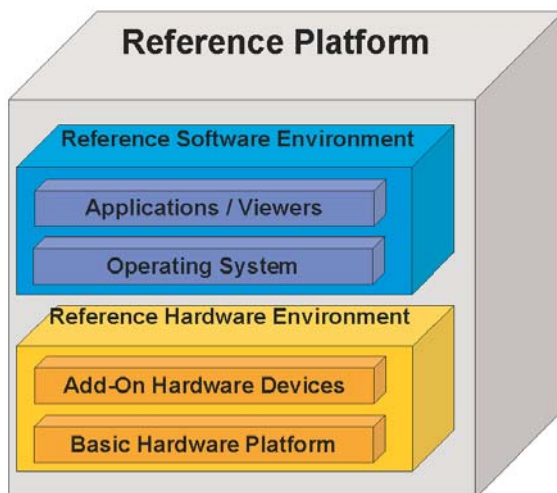


Figure 3.7 / Reference Platform

Reference Software Environment

All the installed software needed to render a set of specific document types or applications:

- ∄ Application / Viewer
Software dedicated to at least support the presentation of a digital object, but potentially to support its creation and modification as well.
- ∄ Operating System
Providing common functionality used by applications and viewers. Also bridging the gap between hardware and software.

Reference Hardware Environment

All the installed physical devices needed to render a set of specific document types or applications

- € Add-On Hardware Devices
All additional devices, e.g. network cards, CD writers, installed in the basic computer.
- € Basic hardware platform
The basic stripped down computer: processor, memory, disks, video, keyboard, sound card.

Initially DIAS will recognize approximately 30 file types. The FileTypeID is used to specify the digital object type. The KB Reference Platform ensures that all these file types will be accessible. The KB Reference Platform can be seen as a physical representation of the technical metadata needed to render the digital objects.

The KB Reference Hardware Environment will use the Reference Software Environment (also referred to as the Disk Image) as a representation of the technical metadata needed by the digital objects stored in DIAS. The Reference Software Platform can be viewed as a specific PLM implementation. Because of the time this solution will be active (± 2 years) only minor changes will be made to the Reference Software Environment during this period and almost none to the Reference Hardware Platform. No major long-term preservation efforts, i.e. migration or emulation, are expected to take place during this brief period.

The introduction of the Preservation Subsystem, expected in 2003, will provide the infrastructure for defining and managing specific View Paths for specific digital object types. More detailed PLMs will then be defined and related to the Disk Image. See Figure 3.8.

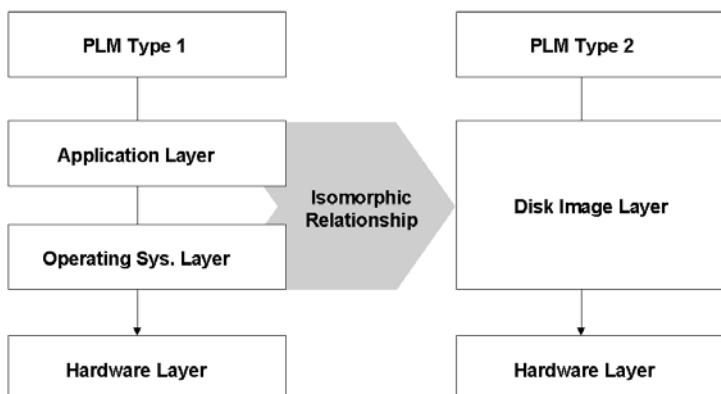


Figure 3.8 / Disk Image PLM associations

This does not automatically imply that the introduction of the Preservation Subsystem will make the use of disk images obsolete. The concept of a Disk Image can speed up the total process for both the ingestion and the delivery processes. Instead of selecting detailed View Paths for each digital object type, most objects can be represented by

specifying one general Disk Image. More detailed technical metadata, using more detailed PLMs and View Paths, can be introduced and associated to the PLM "Disk Image" off-line.

The KB Reference Hardware Environment will use the Reference Software Environment (also referred to as the Disk Image) as a representation of the technical metadata needed by the digital objects stored in DIAS. The Reference Software Platform can be viewed as a specific PLM implementation.

3.5 Installed Electronic Publications

Some of the digital objects (EPublication in DIAS terminology) are not rendered by viewer applications. Within these applications the viewer logic and the actual information content are merged into one indistinguishable digital object. These digital objects have to be installed on a system in order to render their information. The most prominent example is CD-ROMs. They often have to be installed before they can be used.

Within DIAS these types of digital objects actually generate two AIPs:

€ **EPublication**

The raw source, e.g. the bitstream content of the CDROM.

€ **InstalledEPublication**

The disk image (Reference Software Environment) of the installed EPublication on the KB-specific Reference Hardware Environment.

InstalledEPublications are installed on a clean version of the KB Reference Platform after which the bootable disk image is stored as a separate AIP. The advantage of this approach is that the disk image can be preserved by hardware emulation as suggested by Jeff Rothenberg [Rothenberg 2000] for example.

The introduction of the Preservation Subsystem will also introduce other options in which certain layers will contain install scripts that are used to install the EPublication on the fly. They will also be able to assess whether or not the EPublication can be installed on a particular target environment.

A simple PLM model, containing two layers, reflects the current solution: Disk Image Layer and Reference Hardware Environment. Until the introduction of the Preservation Subsystem this will be sufficient because only minor changes in the KB Reference Platform are expected. With the introduction of the Preservation Subsystem more detailed View Paths can be added retroactively. Just as with the normal EPublications maintained in the initial release of the DIAS system.

3.6 Preservation of Technical Metadata

The technical metadata managed by the Preservation Subsystem will be extremely important for defining and managing the long-term preservation activities. They provide the requirements placed on the particular IT infrastructure rendering the digital objects. They are also the main source for defining and executing preservation strategies (migration, emulation) when certain IT infrastructure components become obsolete.

The PLMs and View Paths grow by addition, i.e. there is no modification or deletion. This provides historical information about digital object migration and emulation history. It also raises the quality of the system as a whole because no modifications are allowed. When something goes wrong one can always try to reconstruct the rendering process of a digital object that is no longer accessible from the information provided by the technical metadata history.

Three aspects are relevant to the preservation of the data used by the Preservation Subsystem:

€ **PLM metadata**

The meaning of the concepts such as PLM, Layer and Layer Attributes has to be understood. This will be largely coded within the Preservation Subsystem functionality and the associated documentation, of which this report is one element.

€ **The PLMs and View Paths actually defined**

The PLMs and View Path are only added - never deleted - just like the original AIPs will always appear in DIAS even when they are migrated. All the data used by the Preservation Subsystem will be stored in an RDBMS and thus has access to the standard backup and recovery functionality the RDBMS provides. In addition, the PLM and view path information will be archived into DIAS as a separate AIP on a regular basis.

€ **AIP associations**

Without the associations of the View Paths to the AIPs the technical metadata could not be related to the appropriate digital objects. See Appendix D Figure D.1. The same activities associated with PLMs and View Paths are also performed for this information associating particular View Paths to a FileTypeID in the Table of Contents file of every AIP.

One could argue that archiving the PLM, View Path and AIP association from the RDBMS into an AIP is insufficient by itself. Within the next 100 years the concepts behind the RDBMS could become obsolete. There would not be a system that would be capable of reading in the backup data sets. The UVC data preservation approach [Lorie 2002] could be used to generate a logical data view of the tables stored in the RDBMS. The logical data view can be interpreted without the need for a specific IT infrastructure and therefore will guarantee that the information can always be interpreted in order to be loaded into a future DIAS-like system.

4/

01 preservation 1000010 process model

With the aid of the Preservation Subsystem core presented in Chapter 3 we are now able to take a closer look at the different activities to be conducted by the Preservation Subsystem. The different processes executed by the Subsystem will present a clear picture of the activities related to the long-term preservation of digital objects and the packaging and delivery of digital objects to DIAS users.

We will use a process description methodology called Line of Visibility Enterprise Modeling (LOVEM). LOVEM is a people-oriented IBM methodology using an integrated set of graphics to document, evaluate, and redesign process flows. This report will use LOVEM charts to graphically describe the processes and associated activity flow in which the Preservation Subsystem is involved during daily DIAS operations (Figures 4.1-4.3).

A LOVEM chart is divided into bands, which represent the roles that participate in the execution of a process. Activity flows within the bands are represented by arrows pointing to the next activity to be executed. At the bottom, below the so-called "Manual/Automated" line, the different information systems used during the execution of an activity are listed. Arrows moving to / from information systems represent data movement between the information system and the activity. Finally time is modeled by reading a LOVEM chart from left to right (later in time). Time is not modeled linearly; only the activity order within the process is indicated.

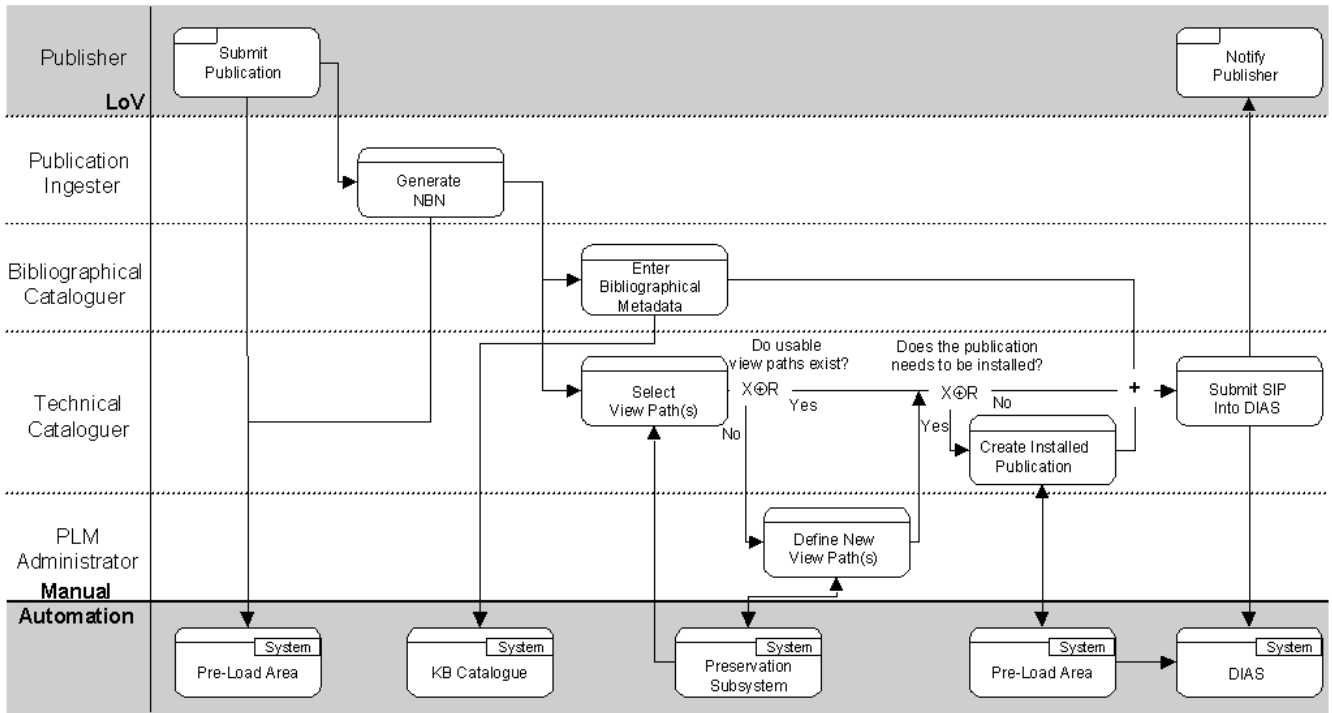
4.1 Publication Ingest

First the digital objects submitted by the publishers or resulting from digitizing projects have to be ingested into DIAS. DIAS uses the concept of an electronic mailroom in which the digital objects are prepared to be ingested by the system. This networked file area will be referred to as the pre-load area.

Figure 4.1 shows the general activity flow for the ingest process for a digital object. In total four roles participate in all identified preservation processes:

€ **Publication Ingester**

Is responsible for checking electronically submitted publications for completeness and for loading them into the pre-load area in preparation for the publication ingest.



X| R (Exclusive Or) denotes that only one or the other of the flows can happen.
 |/ R (Inclusive Or) denotes that one or the other or both flows can happen.

Figure 4.1 / LOVEM chart for publication ingest

- € **Bibliographical Cataloguer**
 Identifies the bibliographical metadata associated with the digital object to be entered in the KB Catalogue to define meaningful search queries on the DIAS content. Both systems can relate objects using the AIP identification ID in the instance of the National Bibliography Number (NBN).
- € **Technical Cataloguer**
 Identifies the technical metadata (View Paths) associated with the digital object. This information will be stored and maintained by the Preservation Subsystem. The link between View Paths and digital object in DIAS is established using file types.
- € **PLM Administrator**
 Defines new View Paths and PLMs when needed either through the introduction of new digital object types or changes in some of the IT infrastructure components.
- € **Preservation Officer**
 Monitors technology changes and how they impact the accessibility of the digital objects stored in DIAS. If digital objects are in danger of becoming inaccessible a preservation strategy has to be defined and executed (migration or emulation) to prevent the digital objects from becoming inaccessible.

Roles and persons do not have to map one to one. A person can play more than one role. The people serving as the PLM Administrator and the Preservation Officer need a more technical / computer science background in order to define new PLMs and View Paths and to assess the impact of technology changes.

Two mainstream activities can be identified during ingest (Figure 4.1): entering bibliographical metadata and entering technical Metadata (View Paths). Within the context of this report we are less interested in the finer grain activity of entering the

bibliographical metadata. The bibliographical metadata is fully controlled by the KB Catalogue system and only interacts with DIAS through its references via an NBN number, which uniquely identifies an AIP.

Estimates show that the registration of the associated technical metadata is a simple matter of selecting an already existing View Path for over 90% of all ingested digital objects. Especially during the initial version of the DIAS (without Preservation Subsystem) only those digital objects will be ingested that can be rendered by the KB Reference Platform, i.e. the Reference Software Environment (Disk Image). But even with the Preservation Subsystem implemented, the initial registration could still be handled using Disk Images. See section 3.4.

Only when no existing View Path is applicable must a new View Path, and potentially even a new PLM, must be defined by the PLM Administrator for that particular digital object type.

If the digital object to be ingested has to be installed, like the Microsoft's Encarta CD-ROM, it first has to be installed on the KB Reference Hardware Platform, after which a Disk Image is produced containing the installed digital object. In this case two AIPs are actually generated:

1. An AIP with the digital object to be originally installed.
2. An AIP with a disk image of the installed digital object on the KB Reference Platform.

Typical questions and activities associated with the Preservation Subsystem in this context are:

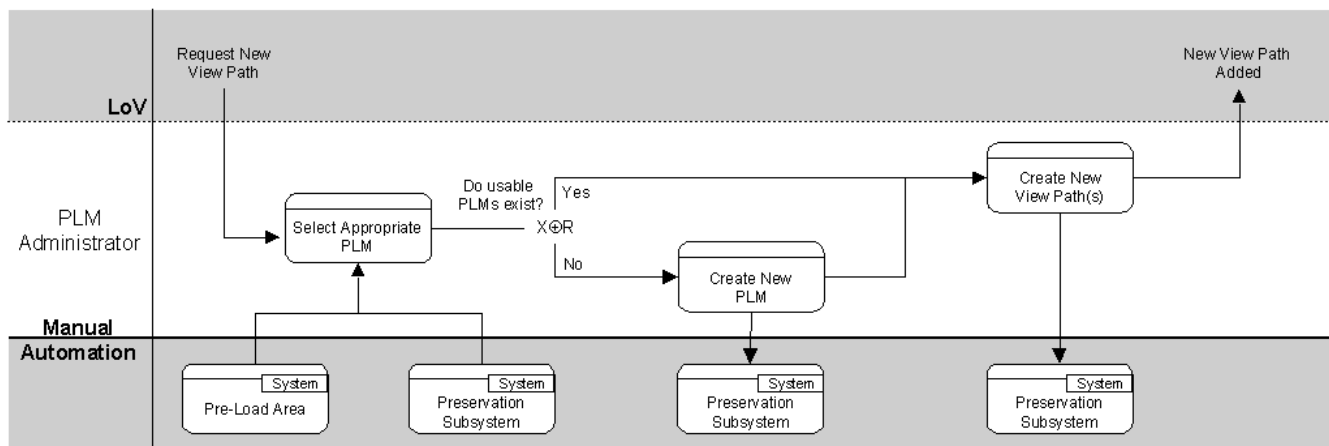
- € Show defined View Paths for specific digital object types.
- € Zoom in on individual layers.
- € Compare different View Paths for the same digital object types.

Only when no existing View Path is applicable must a new View Path and potentially even a new PLM be defined by the PLM Administrator for that particular digital object type.

4.2 Creating new PLMs and View Paths

This process is fully controlled by the PLM Administrator. With the aid of the PLM metadata maintained by the Preservation Subsystem and the already defined View Paths, this person will decide what technical metadata has to be maintained for the new digital object type.

The View Path can either be established on the basis of an existing PLM or a new PLM first has to be defined. See Figure 4.2. Chapter 3 identifies the different components from which a PLM is constructed, i.e. layers and layer attributes. Sometimes the definition of the new PLM will only be a matter of restructuring the layers and adding additional layer attributes. In other cases completely new layers and attributes will have to be defined within the PLM metadata.



X| R (Exclusive Or) denotes that only one or the other of the flows can happen.

| | R (Inclusive Or) denotes that one or the other or both flows can happen.

Figure 4.2 / LOVEM chart for creating new PLMs and View Paths

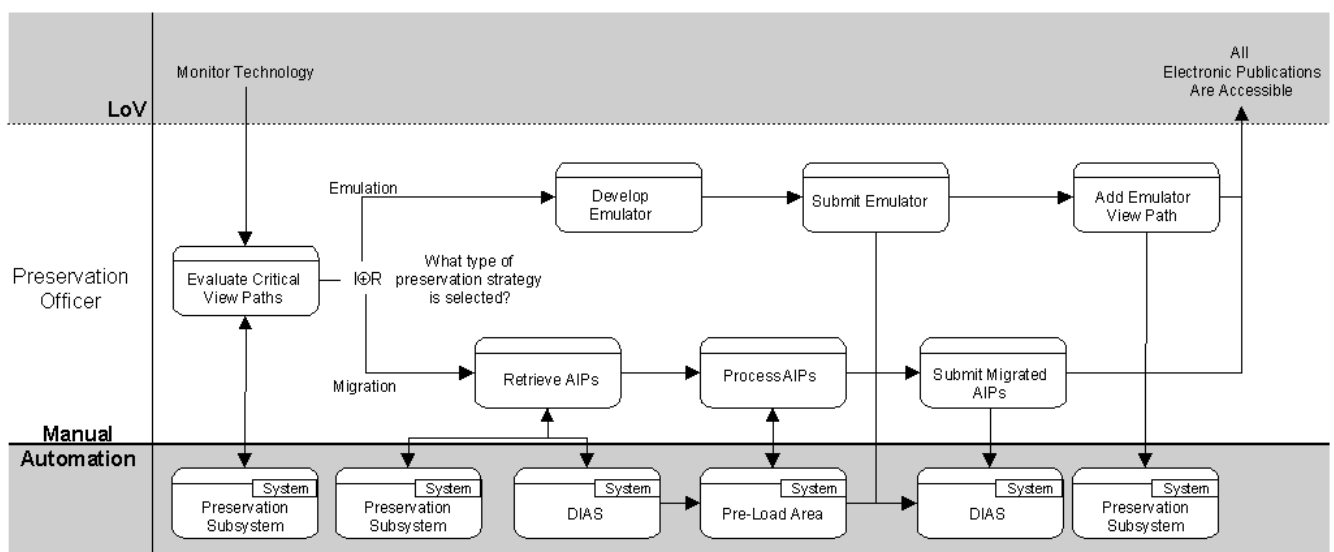
Typical questions and activities associated with the Preservation Subsystem in this context are:

- ∄ Define and store new PLM models.
- ∄ Query regarding the existing layers to be potentially included in a new PLM.
- ∄ Define and store new layer.
- ∄ Query regarding the existing attributes and associated domains to be included in a new layer.
- ∄ Define and store new attributes and domains.

4.3 Monitor Technology

The primary responsibility of the Preservation Officer is to monitor technology changes and their impact on the accessibility of the digital objects stored in the system. When digital objects are in danger of becoming inaccessible a preservation strategy has to be defined (migration or emulation) to prevent the digital objects from becoming inaccessible.

Figure 4.3 shows the basic activities associated with the process of monitoring technology and implementing the preservation strategies.



X| R (Exclusive Or) denotes that only one or the other of the flows can happen.

I| R (Inclusive Or) denotes that one or the other or both flows can happen.

Figure 4.3 / LOVEM chart for monitor technology and execute preservation strategies

The actual implementation of a preservation strategy will result in additional AIP submissions and new entries in the KB Catalogue. However, the original AIP will not be modified. Migrated digital objects will be stored in their own AIP and potential emulation strategies will be defined within the View Paths with an attribute that links back to the emulation program needed, which is stored as a separate AIP.

Typical questions and activities associated with the Preservation Subsystem in this context are:

- ∄ Number of active View Paths associated with the document type.
- ∄ Information on individual layers.
- ∄ The active View Paths associated with the document type which include a specific layer.
- ∄ The active View Paths associated with the document type which include specific attributes.

- € A set of PLMs that depend on a certain layer.
- € Homomorphic relationship between a specific layer of a PLM and other layers of different PLMs.
- € Number of View Path instances associated with a certain PLM.
- € Number of View Path instances associated with a specific layer.
- € Addition of View Paths and PLMs to support the migration/emulation strategy.
- € Maintenance of the associations between AIPs and new View Paths.

The actual implementation of a preservation strategy will result in additional AIP submissions and new entries in the KB Catalogue. However, the original AIP will not be modified. Migrated digital objects will be stored in their own AIP and potential emulation strategies will be defined within the View Paths with an attribute that links back to the emulation program needed, which is stored as a separate AIP.

5/

01 summary



A number of general design principles have guided the implementation of the Preservation Subsystem:

€ **Flexibility**

Functionality should be implemented as flexible and general as possible to provide a sound foundation for the required changes in PLMs over time.

€ **Loosely coupled**

Functionality should be loosely coupled reducing the dependency among the different components, e.g. preservation, storage, catalogue. There should only be limited links between different components of DIAS, e.g. NBN, FileTypeID.

€ **Redundancy without complexity**

Redundancy should be implemented but this should only minimally increase the complexity. We elected not to include the AIP Table of Contents in the Preservation Subsystem because this would have made the dependencies and interaction between the Ingestion and Preservation module more complex without providing major added benefits.

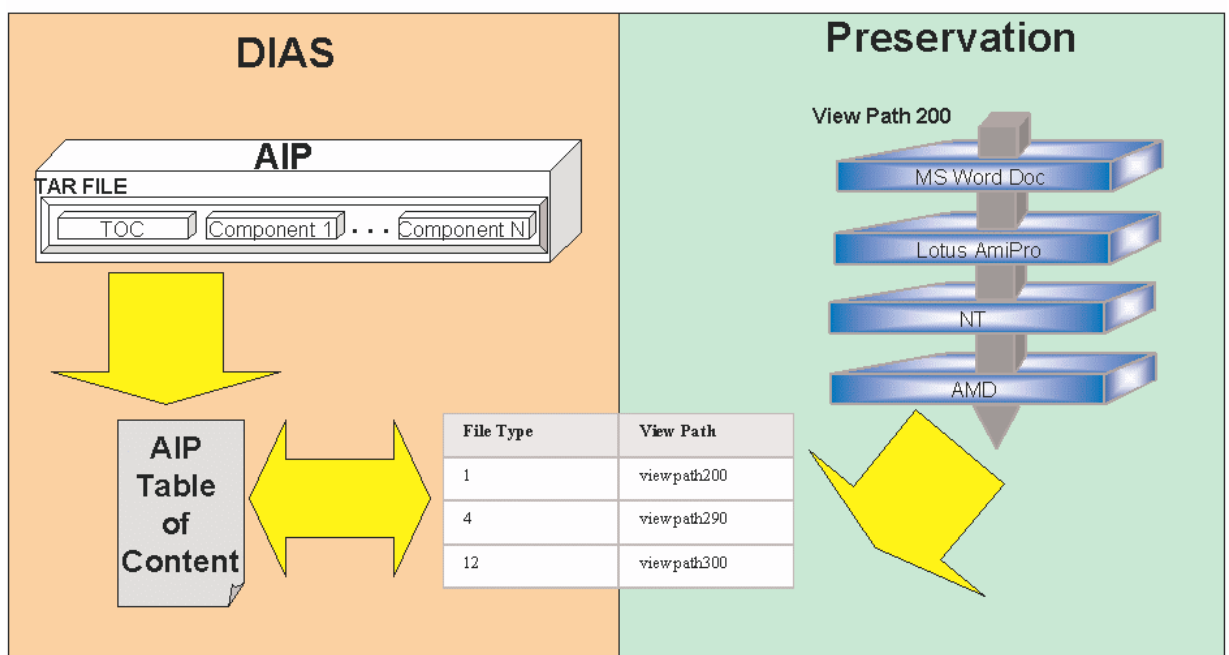


Figure 5.1 / DIAS overview - preservation interaction

Figure 5.1 provides an overview of the complete process of linking technical metadata to digital objects in the AIP.

The separation of the technical metadata from the AIP eliminates the need for AIP and View Path modification. This reduces the chances of making mistakes or introducing corruption due to repeated modification.

In summary, the following preservation functionality objectives and requirements have been identified:

The preservation functionality has two objectives

- € Activities associated with technical preservation, i.e. implementation of migration and emulation strategies.
- € Generating the requisite environments during electronic object delivery.

Multiple PLMs are needed to define usable View Paths

- € A metadata model is used to define multiple PLMs.
- € View paths are linked to AIPs by having the Table of Contents register an associated file type for each file included in the AIP.
- € The technical metadata (PLMs, View Paths) is regularly archived as an AIP within the DIAS.
- € The PLMs and View Paths grow by addition, i.e. there is no modification or deletion.

AIPs can contain multiple components

- € AIPs are stored and are not subsequently modified.
- € The Table of Contents includes the file type identifications.

Due to a lack of clarity with regard to the Preservation Subsystem requirements, the Subsystem was not included in the first version of DIAS. In order to still create a workable system and prepare for the introduction of the Preservation Subsystem, the following decisions have been made to provide the current limited support for preservation in the system:

Assumptions for the next 2 years

- € There will only be a limited number of Reference Platforms;
- € The selected Reference Platforms will continue to be active over the two-year period.
- € Current simple digital object types (i.e. not installed) do not require the support of a View Path to remain accessible, i.e. is to be provided by the KB Reference Platform.

What we proposed

- € Simple digital objects register the associated identified file types using an association with the AIP and are indexed at a component level in the Table of Contents;
- € To be installed digital objects are preserved for the short term by archiving the installed version on the KB Reference Platform and are identified as such by their file type;
- € The PLM models associated with digital objects will vary over time with the increased experience gained. Thus, we propose starting this process once the preservation functionality is implemented;
- € The AIP file type association will guarantee connectivity between the DIAS system and the Preservation Subsystem.

01 appendix a: 01000010 references

[Booch et al. 1999]

Booch, G., Rumbaugh, J., Jacobson, I., *The Unified Modeling Language User Guide*, Addison-Wesley, Reading, MA, 1999.

[CCSDS 2001]

Management Council of the Consultative Committee for Space Data Systems, *CCSDS650.0-R-2: Reference Model for an Open Archival Information System (OAIS)*. Red Book, Washington, DC, July 2001, http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html.

[Diessen and Werf-Daselaar 2002]

Diessen, R.J. van and Werf-Davelaar, T. van der, *Authenticity in a Digital Environment*, KB/IBM Long-Term Preservation Study Report Series Number 2, December 2002.

[Lorie 2002]

Lorie, R., *The UVC: a Method for Preserving Digital Documents - Proof of Concept*, IBM / KB Long-Term Preservation Study Report Series Number 4, December 2002.

[Lupovici and Masanès]

Lupovici, Catharine, and Julien Masanès, *Metadata for the Long-Term Preservation of Electronic Publications*, NEDLIB Report Series Number 2, September 2000.

[Rothenberg 2000]

Rothenberg, J., *An Experiment in Using Emulation to Preserve Digital Publications*, NEDLIB Report Series, April 2000.

[Tanenbaum 1990]

Tanenbaum, A.S., *Structured Computer Organization*, Prentice-Hall, Engle wood Cliffs, N.J., 1990.

01 appendix b: 01000010 glossary

Archival Information Package (AIP): Content Information and the associated Preservation Description Information required to preserve the Content Information over the long term. This information includes the related Packaging Information.

Archival Storage: The OAIS entity that contains the services and functions used for the storage and retrieval of Archival Information Packages.

Content Information: That set of information that is the primary target for preservation. This is composed of a Data Object and its Representation Information. An example of Content Information could be a single table of numbers representing, and understandable as, temperatures, but excluding the documentation that would explain its history and origin, how it relates to other observations, etc.

Data Management: The OAIS entity that contains the services and functions for populating, maintaining, and querying a wide variety of information such as catalogues and inventories of what may be retrieved from Archival Storage, processing algorithms that may be run on retrieved data (if any), consumer access statistics, security controls, and OAIS schedules, policies, and procedures.

Digital Information Archiving System (DIAS) is the core of the KB's electronic deposit system. Version 1 has been developed by IBM and was released in October 2002.

Digital Object: An object composed of a set of bit sequences.

Dissemination Information Package (DIP): An Information Package that contains part or all of one or more AIPs and that is distributed to the consumer as requested.

DNEP: In September 2000 the KB and IBM Netherlands signed the final contract which initiated the project "Depot voor Nederlandse Electronische Publicaties" (DNEP) [Deposit for Dutch Electronic Publications] to design and implement DIAS with a Long-Term Preservation Study as an integral part of the total effort.

Information often is incorrectly used to refer to data, but information is synonymous with information process, see information process.

Information Package: Content Information and associated Preservation Description Information that is needed to aid in the preservation of the Content Information. The Information Package has Packaging Information associated with it, which is used to delimit and identify the Content Information and Preservation Description Information.

Information Process is the interpretation of observed phenomena and data in a particular context at a particular time by a person or a group.

Ingest: The OAIS entity that contains the services and functions that accept Submission Information Packages from Producers, prepares Archival Information Packages for storage, and ensures that Archival Information Packages and their supporting Descriptive Information are included in the OAIS.

KB: The National Library of the Netherlands (Koninklijke Bibliotheek, KB).

Logical view of the data: a view of the data that is easily understandable because it follows the way the user normally thinks about the data, rather than the internal representation often designed for efficiency.

Long-Term Preservation: The act of maintaining information, in a correct and independently usable form, over the long term. Independently usable information has sufficient documentation to allow the information to be understood and used by the designated community without having to resort to special resources not widely available, including named individuals.

Metadata: Data about other data.

Migration: The transfer of digital information within the DIAS with the intention of preserving this information. In general this is distinguished from transfers by three attributes:

- € The focus is on preserving the full information content.
- € The newly archived information is intended to replace the old archive.
- € It is understood that DIAS has full control over and responsibility for all aspects of the transfer.

National Bibliography Number (NBN): This is a generic name referring to a group of identifier systems utilized by the national libraries and only by them to identify deposited publications which do not have an identifier, or to identify descriptive metadata (cataloging) that describes the resources.

Open Archival Information System (OAIS): OAIS is a functional reference model. An OAIS is an archive consisting of an organization of people and systems that has accepted the responsibility to preserve information and make it available for a designated community. It specifies a specific set of responsibilities and it allows an OAIS archive to be distinguished from other uses of the term archive. The term 'Open' in OAIS is used to imply that this recommendation, as well as future related recommendations and standards are developed in open forums. It does not imply that access to the archive is unrestricted.

Preservation Description Information (PDI): Information that is required for adequate preservation of the Content Information.

Submission Information Package (SIP): The Information Package identified by the producer in the submission agreement with the OAIS.

Universal Virtual Computer or UVC is a virtual machine specially designed to develop programs today that will be able to run on a future machine by simply writing an emulator of the UVC in the future.

XML: Extensible Markup Language.

01 appendix c: 1000010

DIAS release 1

recognized file types

FileTypeID	FileTypeExtension	FileType	FileTypeVersion	MimeType
1	pdf	Adobe Acrobat Document	1.2	application/pdf
2	pdf	Adobe Acrobat Document	1.3	application/pdf
3	bmp	Bitmap Image		image/bmp
4	gif	GIF Image	87a	image/gif
5	gif	GIF Image	89a	image/gif
6	jpg	JPEG Image		image/jpeg
7	jpeg	JPEG Image		image/jpeg
8	jpe	JPEG Image		image/jpeg
9	tiff	TIFImage Document		image/tiff
10	tif	TIFImage Document		image/tiff
11	png	PNG Image		image/x-png
12	zip	PKZIP File		application/zip
13	tar	TAR File		application/x-tar
14	exe	Executable File		application/octet-stream
15	ps	Postscript		application/postscript
16	mpg	Movie Clip		video/mpeg
17	mpeg	Movie Clip		video/mpeg
18	mpe	Movie Clip		video/mpeg
19	avi	Video Clip		video/x-msvideo
20	txt	TXT File		text/plain
21	htm	Hypertext Document	2.0	text/html
22	html	Hypertext Document	2.0	text/html
23	htm	Hypertext Document	3.2	text/html
24	html	Hypertext Document	3.2	text/html
25	htm	Hypertext Document	4.0	text/html
26	html	Hypertext Document	4.0	text/html
27	htm	Hypertext Document	4.01	text/html
28	html	Hypertext Document	4.01	text/html
29	pqi	PQI Image File		Application/octet-stream
30	css	Cascading Style Sheet Document		Text/css

01 appendix d: possible PLM implementation

00010

The proposed implementation is based on the support of a relational database management system. It does not take into account possible design changes made to the tables for performance reasons. The actual View Paths are managed by a series of tables for which the ViewPathTable is the main entry point, see Figure D.1. The AIP Reference table manages the links between the View Paths and a specific FileTypeID and indirectly to the AIPs by the ToC contained in the each AIP. In this case a separate table called AttributeIDValueTable models each attribute type used in a View Path.

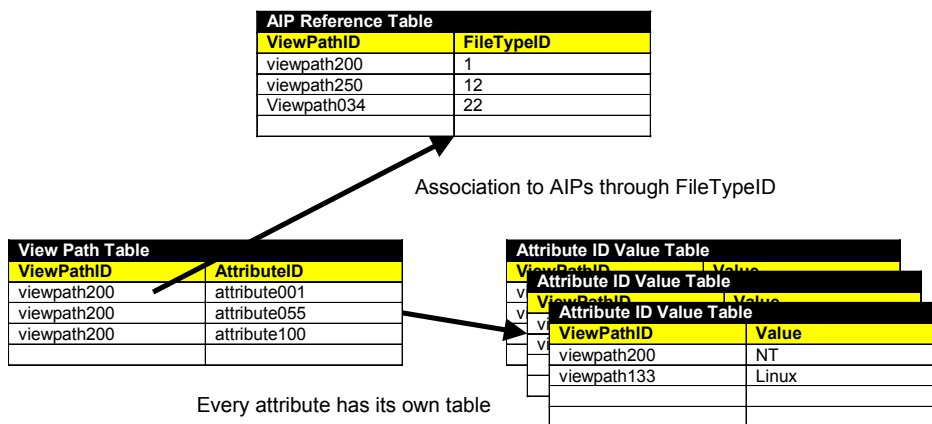


Figure D.1 /View Path implementation in DBMS environment

Another possible implementation for registering the attribute values would be to provide a table for each PLM type with all the attributes in one table, see Table D.1.

PLM Value Table			
ViewPathID	Operating System	Application	Version
viewpath178	NT	Acrobat	3.0
viewpath345	Windows2000	Netscape	7.0

Table D.1 One tabel for each PLM

The metadata for defining the PLMs based on Layers and Attributes can be represented by the table structure shown in Figure D.2.

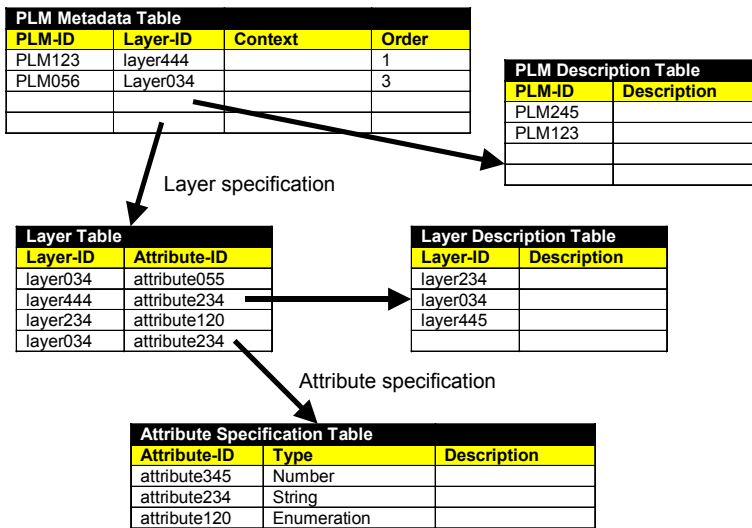


Figure D.2 / PLM metadata tables

01 appendix e: 01000010 Use Cases

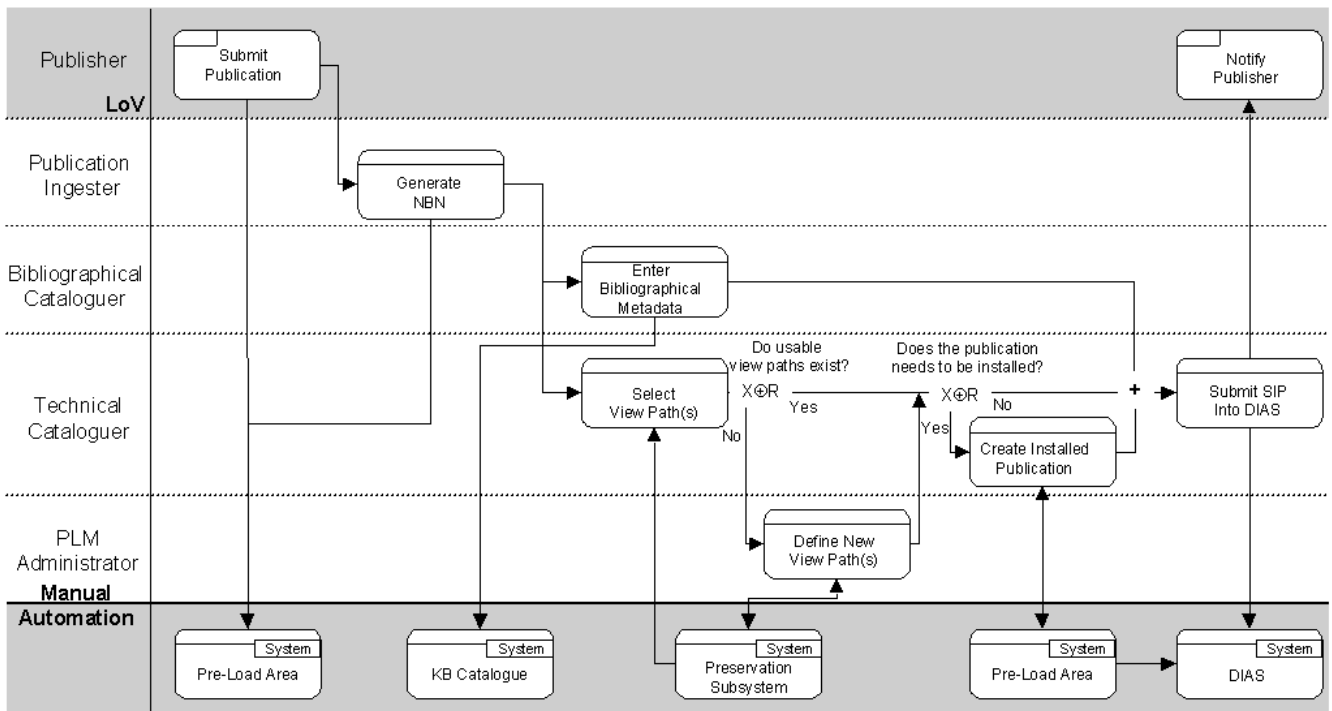
Actor Name	Publication Ingestor
Brief Description	The person responsible for preparing a digital object (or objects) for ingestion into the DIAS system.
Associations To Use Cases	001- Create NBN.

Actor Name	Bibliographical Cataloguer
Brief Description	The person responsible for entering all the relevant bibliographical metadata associated with the digital object(s) in the KB catalog system.
Associations To Use Cases	002- Enter bibliographical metadata.

Actor Name	Technical Cataloguer
Brief Description	The person responsible for selecting the appropriate View Paths associated with the digital object(s). He/she also creates an Installed EPublication from the digital object(s) when necessary.
Associations To Use Cases	003- Select View Path(s), 005- Create installed publication, 006- Ingest the electronic publication into DIAS.

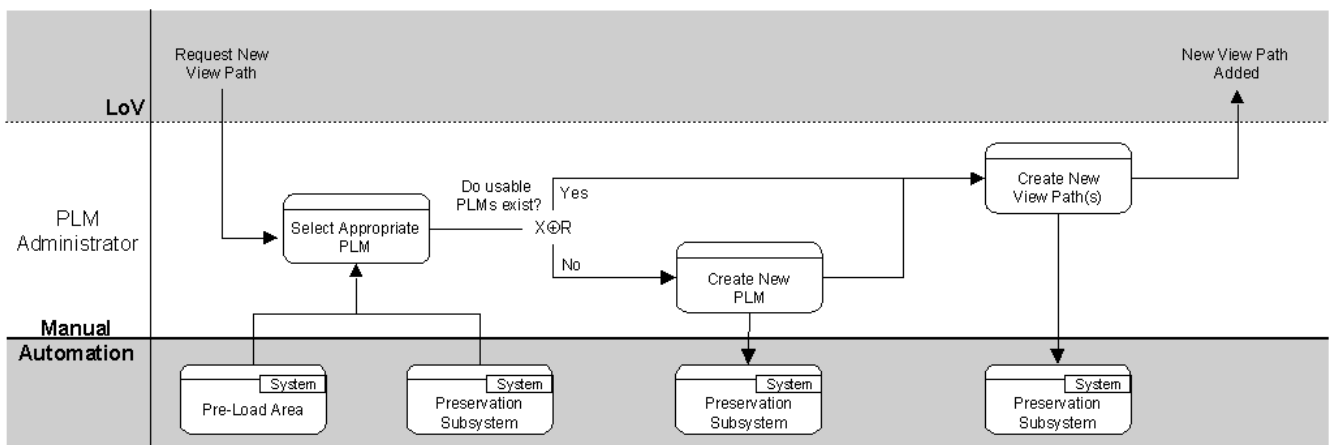
Actor Name	PLM Administrator
Brief Description	Person who defines new View Paths and PLMs when needed either by introducing new digital object types or making changes in some of the IT infrastructure components.
Associations To Use Cases	004- Define View Path, 007- Select appropriate PLM, 008-Create new PLM.

Actor Name	Preservation Officer
Brief Description	Person who monitors technology changes and their impact on the accessibility of the digital objects stored in the DIAS system. If digital objects are in danger of becoming inaccessible this person has to define a preservation strategy (migration or emulation) to prevent these digital objects from becoming inaccessible.
Associations To Use Cases	009- Evaluate critical View Paths, 010- Develop Emulator, 011- Submit Emulator, 012- Add Emulator View Path, 013- Retrieve AIPs, 014- Process AIPs, 015- Submit Migrated AIPs.



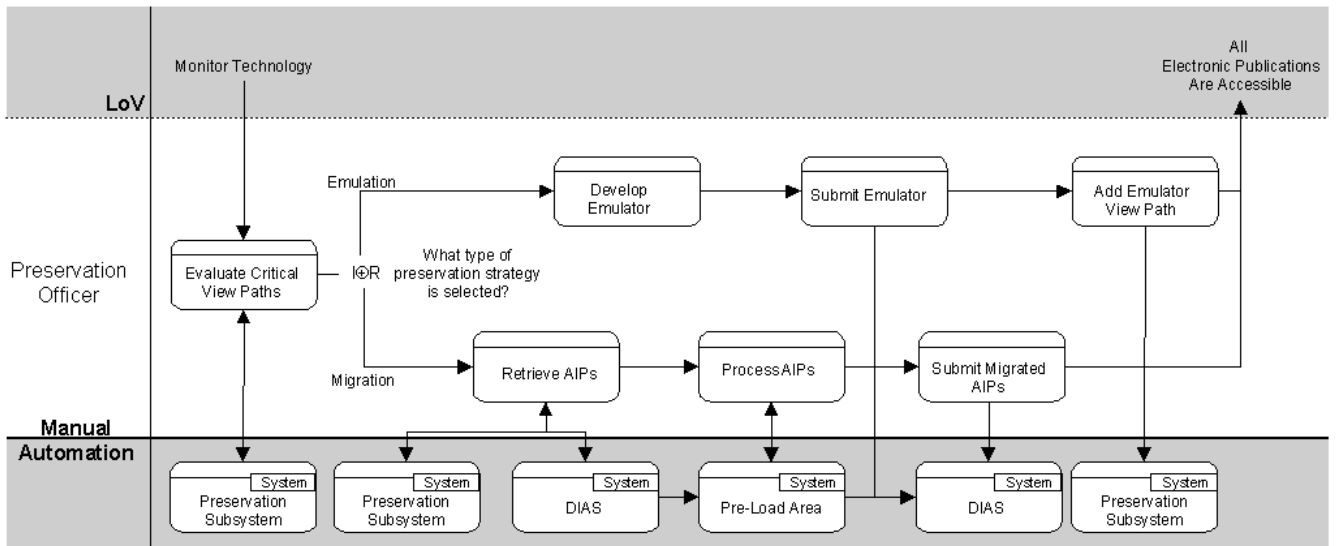
X⊕ R (Exclusive Or) denotes that only one or the other of the flows can happen.
 ⊕ R (Inclusive Or) denotes that one or the other or both flows can happen.

Figure E.1 / LOVEM chart for publication Ingest



X⊕ R (Exclusive Or) denotes that only one or the other of the flows can happen.
 ⊕ R (Inclusive Or) denotes that one or the other or both flows can happen.

Figure E.2 / LOVEM chart for creating new PLMs and View Paths



X| R (Exclusive Or) denotes that only one or the other of the flows can happen.
 I| R (Inclusive Or) denotes that one or the other or both flows can happen.

Figure E.3 / LOVEM chart for monitor technology and execute preservation strategies

Activity #001	Create NBN
Subject Area	Pre-Load Area
Business Event	Submit electronic publication
Actor(s)	Ingester, Pre-load System
Overview	<p>The NBN is the National Bibliography Number that is the external identifier of each (digital) asset of the KB. The NBN is obtained from a NBN generator (out of scope). The NBN syntax conforms to a Uniform Resource Name (URN).</p> <p>NBN ::= "urn:" + <NID> + ":" + <NSS></p> <p>NID stands for Namespace Identifier and NSS stands for Namespace Specific String. In the case of a NBN URN the NID will be "NBN". The NSS will have the following syntax:</p> <p>NSS ::= <country-code> + ":" + <sub-name-space> + "-" + < nbn-string></p> <p>The country-code for the KB will be "nl", the sub-name-space for DIAS. The nbn-string is a string of at least 1 and a maximum of 255 characters. These characters must comply with the NSS syntax as defined in RFC2141 (see http://www.ietf.org). See also RFC2288.</p> <p>An example of a valid NBN for KB will be:</p> <p>NBN ::= "URN:NBN:nl:kb:eDepot" + <nbn-string></p> <p>DIAS will only store the nbn-string part of the NBN. This nbn-string will also be used to retrieve AIPs from DIAS.</p> <p>DIAS will use the name NBN for the nbn-string.</p>
Precondition	Either the electronic publication has been received off-line (tape or CDROM) or on-line (ftp, http or email). The digital asset has been copied to the pre-load area to be tested for completeness and correctness.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. The electronic publication is complete and has been tagged with an NBN.	The electronic publication is complete and NBN has not yet been used in DIAS.
2. DIAS rejects the electronic publication for ingestion.	Either the electronic publication has been deemed to be incomplete or the suggested NBN is already being used in DIAS.
Description of Termination outcome #1	The electronic publication has been tagged with its unique NBN and stays in the pre-load area for further processing.
Business Rules	The origin of the electronic publication has to be a recognized publisher with which the KB has an agreement to incorporate the publications in its electronic deposit.
Input	Generated NBN, electronic publication.
Output	NBN tagged electronic publication.

Activity #002	Enter bibliographical metadata
Subject Area	Collection Management
Business Event	Add bibliographical metadata
Actor(s)	Bibliographical Cataloguer, Catalogue System
Overview	An entry in the KB's bibliographical metadata system will be added for the electronic publication based on the content of the electronic publication and optional accompanying metadata supplied by the publisher.
Precondition	The publication has been tagged with an NBN indicating it has been checked for completeness and correctness.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. All relevant bibliographical metadata has been entered in the Catalogue System.	The bibliographical metadata needed to manage KB collections has been supplied with the electronic publication.
2. No entry in the Catalogue System	The ingestion process will be placed on hold until the relevant bibliographical metadata can be supplied by the publisher.
Description of Termination outcome #1	The Catalogue System already contains the bibliographical metadata needed to access the electronic publication.
Business Rules	Bibliographical metadata identified by the NBN have to be present before the electronic publication can be archived.
Input	NBN, electronic publication.
Output	Entry in the KB Catalogue system for the electronic publication.

Activity #003	Select View Path(s)
Subject Area	Preservation Management
Business Event	Add technical metadata
Actor(s)	Technical Cataloguer, Preservation Subsystem
Overview	The first activity is to evaluate the exact technical infrastructure needed to render the electronic publication. In most cases the format(s) used in the electronic publication will already be known and View Paths will already exist.
Precondition	The publication has been tagged with an NBN indicating that it has been checked for completeness and correctness.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. All file formats included in the electronic publication are linked to a defined View Path in the Preservation Subsystem.	A complete set of View Paths for this electronic publication has already been defined in the Preservation Subsystem.
2. Additional View Paths have to be defined before the electronic publication can be ingested into DIAS.	The electronic publication submitted is of a new type or of a higher version than previous submitted electronic publications in this category.
Description of Termination outcome #1	All the file formats used in the electronic publication are represented by an active View Path in the Preservation Subsystem and thus can be rendered on a specific current reference workstation.
Business Rules	All format types encountered in electronic publications have to have at least one active View Path to make sure the electronic publication can be rendered.
Input	List of all the file types encountered in the electronic publication.
Output	List of selected View Paths.

Activity #004	Create View Path
Subject Area	Preservation Management
Business Event	A non-supported file type and/or version encountered
Actor(s)	PLM Administrator, Preservation Subsystem
Overview	The type of software configuration (operating system and applications) needed to render the particular document on the reference workstation will be evaluated. In many cases View Paths known already can be used, for instance Microsoft Word which can both read .doc files and RTF files. In other cases additional software has to be installed on the reference platform and tested to see whether it can render the specific file format under investigation. Then a new View Path will be added to the Preservation Subsystem potentially preceded by the definition of a new PLM.
Precondition	General available application software has to exist to render any specific format type.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. All file types included in the electronic publication are linked to a defined View Path in the Preservation Subsystem	A specific Software Environment can be installed on the Reference Hardware that correctly renders the electronic publication.
Description of Termination outcome #1	In most cases a new View Path can be defined based on one of the existing PLMs already defined in the Preservation Subsystem. Although guided by the PLM the PLM Administrator still needs enough IT knowledge to specify the required technical metadata needed to render and manage the electronic publications stored inside DIAS.
Business Rules	All file types encountered in electronic publications have to have at least one active View Path to make sure the electronic publication can be rendered.
Input	List of all the unknown file types encountered in the electronic publication.
Output	Additional defined View Paths for the specific file types.

Activity #005	Create installed publication
Subject Area	Pre-load Area
Business Event	None
Actor(s)	Technical Cataloguer
Overview	On the reference workstation the electronic publication, e.g. CDROM, will be installed and the disk image will be prepared in the pre-load area to be archived as a separate AIP in DIAS.
Precondition	None
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. The disk image for the installed publication prepared as a separate SIP.	The installation on the reference workstation of the KB was successful.
Description of Termination outcome #1	The Technical Cataloguer installs the publication on a reference workstation, reboots the reference workstation with the standard reference platform and creates a disk image backup of the installed publication. The image backup must be assigned the image backup file type. The hardware and operating system of the reference platform must be selected. The License information has to be archived by including a scanned image as a license sheet. If the publication can be installed on more than one platform the steps are repeated for the other reference platforms.
Business Rules	None.
Input	Installable electronic publication.
Output	The disk image of the installed publication on the reference workstation.

Activity #006	Ingest the electronic publication into DIAS
Subject Area	DIAS Ingestion
Business Event	SIP is ready to be ingested
Actor(s)	Technical Cataloguer, Pre-Load Area, DIAS
Overview	The electronic publication has now been checked and all the technical metadata needed to render the electronic publication is present in the Preservation Subsystem. Everything is packaged into one or more SIPs (in the case of an installed publication) to be ingested by DIAS that will result in a stored AIP for each SIP.
Precondition	SIPs are complete and correct.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. Store AIP in DIAS for each SIP submitted.	The completeness and correctness of the submitted SIPs.
Description of Termination outcome #1	Stored AIP in DIAS for each SIP submitted.
Business Rules	None.
Input	The SIPs.
Output	None.

Activity #007	Select appropriate PLM
Subject Area	Preservation Management
Business Event	Request new View Path
Actor(s)	PLM Administrator, Preservation Subsystem, Pre-Load Area
Overview	When a new file type is being introduced into DIAS at least one new View Path has to be added to specify the technical metadata needed to render the file type. The View Paths will be defined as much as possible according to one of the already defined PLMs. However, in some cases first a new PLM has to be identified in order to register all relevant technical metadata in the View Path.
Precondition	None.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. A usable PLM has been selected to define the View Path.	New file formats have to be evaluated against the authenticity requirements in order to define the needed technical metadata.
2. No usable PLM could be selected to define the View Path.	The complete set of PLM layers and attributes already defined in the Preservation Subsystem are assessed for potential reuse in the newly to be defined PLM.
Description of Termination outcome #1	The selected PLM will guide the PLM Administrator to define the technical metadata to be associated with the specific file type (FileTypeID).
Business Rules	All defined View Paths are based on an existing PLM.
Input	None.
Output	None.

Activity #008	Create new PLM
Subject Area	Preservation Management
Business Event	None
Actor(s)	PLM Administrator, Preservation Subsystem, Pre-Load Area
Overview	A new PLM is added when none of the existing PLMs defined in the Preservation Subsystem can be used as a template to specify the View Path for a specific file type.
Precondition	No existing PLM satisfies the technical metadata needs of the new file format.
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. Added PLM to support the definition of View Paths related to the file type under investigation.	The hardware and software needed to render the file type has to be generally available and no encryption or authorization functionality is embedded.
Description of Termination outcome #1	The complete set of PLM layers and attributes already defined in the Preservation Subsystem are assessed for potential reuse in the newly to be defined PLM. Only as a last resort new technical metadata attributes and layers are defined. The newly defined PLM will be directly available as a template to define new View Paths based on the specific PLM.
Business Rules	None.
Input	PLM, PLM layers, PLM attributes defined in the Preservation Subsystem
Output	None.

Activity #009	Evaluate critical View Paths
Subject Area	Preservation Management
Business Event	Technology Innovations
Actor(s)	Preservation Officer, Preservation Subsystem
Overview	Based on technology innovations and obsolescence the existing View Paths are regularly queried to evaluate whether all electronic publications can still be rendered.
Precondition	None
Termination Outcomes	
Condition Affecting Termination Outcomes	
1. Set of endangered FileTypeIDs.	The queried technical metadata items have to be defined within the Preservation Subsystem.
Description of Termination outcome #1	The View Paths are queried on a PLM attribute level in the Preservation Subsystem to determine the set of View Paths and file types affected. When the number of active View Paths not affected by the queried technical metadata attributes is lower than a by the electronic deposit defined minimum some type of preservation strategy has to be performed.
Business Rules	The number of active View Paths must not drop below a by the electronic deposit defined minimum threshold and never be zero. Zero represents the case in which an object of the specific file type can no longer be rendered by any combination of currently available hardware and software environment.
Input	Complete set of defined View Paths, the PLM attribute level item defined not to be available anymore.
Output	Set of FileTypeIDs for which the number of View Paths would drop under the declared minimum.

Activities #010-#012	Develop Emulator – Submit Emulator – Add Emulator View Path	
Subject Area	Preservation Management	
Business Event	None	
Actor(s)	Preservation Officer, Preservation Subsystem, DIAS	
Overview	One way to reactivate a View Path endangered of becoming obsolete by technology innovations is to develop an emulator for the old environment. This emulator itself has to be stored in DIAS as a separate AIP. Most emulation approaches currently defined will try to emulate the hardware in order to run the original software.	
Precondition	None.	
Termination Outcomes		Condition Affecting Termination Outcomes
1. Emulator developed and tested		The usage of specialized hardware will increase the emulation complexity.
Description of Termination outcome #1	The developed emulator itself has to be stored in DIAS as a separate AIP. In addition a new View Path will be associated with the fileTypeID that describes the technical metadata related to the positioning and usage of the emulator.	
Business Rules	None.	
Input	FileTypeID, associated View Paths.	
Output	Archived emulator to allow rendering of the file type.	

Activities #013-#015	Retrieve AIPs – Process AIPs –Submit Migrated AIPs	
Subject Area	Preservation Management	
Business Event	None	
Actor(s)	Preservation Officer, Preservation Subsystem, DIAS	
Overview	The effected AIPs that include files of the specific set of FileTypeID are retrieved and converted into a better supported file type in the current technological environment.	
Precondition	None.	
Termination Outcomes		Condition Affecting Termination Outcomes
1. All endangered files of the specific FileTypeID have been migrated to a new format.		Special care has to be taken to make sure the authenticity of the original electronic documents is maintained during the migration to a new file type.
Description of Termination outcome #1	<p>The effected AIPs that include files of the specific set of FileTypeID are retrieved and stored in a migration workspace of the Pre-Load Area. There they will be further processed by the Preservation Subsystem and converted to the selected destination FileTypeID after which they are submitted as SIPs into DIAS. We assumed that the destination FileTypeID already has been defined in the Preservation Subsystem together with the associated View Paths.</p> <p>The actual migration of the AIPs will be automated as much as possible. This process will be controlled by the Preservation Subsystem.</p>	
Business Rules	None.	
Input	FileTypeID (file format) endangered by technology innovations.	
Output	None.	

IBM
long

KB
term

preservation
study

